응용통계연구 (2010)
23(4), 705–713

DLBCL 환자의 대사경로 정보를 이용한 생존예측

이광현[1] · 이선호[2]

[1]세종대학교 응용통계학 전공, [2]세종대학교 응용통계학 전공

(2010년 4월 접수, 2010년 5월 채택)

요 약

마이크로어레이 실험 결과로부터 생존예측지표를 개발하는 일은 관찰 유전자수가 환자의 수보다 훨씬 많고 또 반응변수가 중도절단이 포함된 생존시간이기 때문에 어려운 작업이다. 또한 개별유전자 분석의 문제점이 대두되면서 동일한 대사기능을 수행하는 유전자들의 집합을 대상으로 분석하는 방법이 대두되고 있다. DLBCL 환자들의 마이크로어레이 유전자 발현 자료와 생존시간, 유전자들의 대사경로 정보를 바탕으로 생물학적 해석이 쉬운 생존예측지표를 찾고 그 정확성을 검정하는 pilot study를 실시하였다. 또한 유전자 걸러내기가 지표의 효율성에 미치는 영향력도 비교하여 보았다.

주요용어: 마이크로어레이 실험, 생존분석, 대사경로, 주성분분석, 비례위험모형.

1. 서론

동시에 많은 유전자의 발현 상황을 총체적으로 탐색가능하게 한 마이크로어레이 기술의 발전은 질병의 조기진단과 치료를 위한 바이오마커를 찾아내고 신약 개발에 큰 기여를 하고 있다. 또한 마이크로어레이 자료로부터 종양환자들의 전이여부와 생존기간을 예측할 수 있는 지표의 개발은 환자의 개인별 맞춤치료를 가능하게 하고 삶의 질도 향상시킬 수 있다. 그러나 마이크로어레이 실험은 표본의 수(n)가 유전자 수(p)에 비해 극히 작은 문제(small 'n', large 'p')가 발생하여 일반적인 통계분석법 적용에 어려움이 많다. 또한 수만개 유전자들의 개별 발현값을 기초로 하기 때문에 연구결과의 생물학적 해석이 어렵고 동일 질병이라도 다른 자료집합을 사용하면 분석 결과들 사이에 공통점을 찾기 힘들다는 문제점(Subramanian 등, 2005)이 제기되었다. 이런 점을 개선하기 위하여 생물학적 연구를 통해 구축된 데이터베이스의 정보를 이용하여 기능이나 대사작용이 비슷한 유전자들끼리 묶어 연구하는 유전자 집합 분석이 대두되었고 환자의 표현형(phenotype, 즉 종양의 종류, 암전이 여부, 생존 기간 등)에 따라서 유의한 차이를 나타내는 대사경로(pathway)를 찾아내는 유의성 검정에 대한 연구가 활발히 진행되고 있다.

마이크로어레이 자료로부터 개별유전자 분석을 통해 이진표현형(종양/정상, 전이 여부 등)에 대한 지표를 개발하는 연구는 많이 진행되었으나 표현형이 생존시간인 경우는 축적된 자료가 많지 않으며 중도절단의 문제가 있어 활발히 다루어지지 않았다. 또한 유전자 각각의 발현자료 뿐만 아니라 데이터베이스에 축적된 그들의 생물학적인 정보도 이용해 지표를 개발하는 방법에 대한 연구는 아직도 초기단계이다.

본 논문에서는 미만성 거대B세포 림프종(diffuse large-B-cell lymphoma; DLBCL) 환자들을 장기간 추적한 생존시간과 그들의 마이크로어레이 자료로부터 생존예측지표를 생성하고 검증하려 하는데 대사

이 논문은 2008년 교육인적자원부의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(2008-531-C00019).

[2]교신저자: (143-747) 서울시 광진구 군자동 98, 세종대학교 응용통계학 전공, 교수. E-mail: leesh@sejong.ac.kr

경로 분석에 기초한 결과는 종양에 내재된 메카니즘을 반영하며 데이터의 차이에 민감하지 않고 생물학적 해석이 쉬워지리라 기대한다.

2장에서는 대사경로 정보를 이용하여 지표개발에 필요한 과정들을 단계별로 나누어 설명하고 3장에서는 실제자료를 이용하여 생존예측지표를 추정하고 정확성을 검정하였다.

2. 지표개발에 필요한 과정에 대한 고찰

종양 환자들은 수술 후 재발 가능성을 낮추기 위해 항암요법 치료를 추가로 시행하는 경우가 많다. 그러나 초기 유방암 진단을 받은 환자들의 유전자 발현 자료로부터 찾아낸 전이예측지표는 수술 후 항암치료 없이 타목시펜(tamoxifen) 투약만으로 암의 재발 방지 효과를 볼 수 있는 환자들을 가려내는 역할을 하며 이것을 새로운 환자 20명에게 적용한 결과 16명에 대한 판단이 정확하였다는 연구결과가 있다 (Ma 등, 2004). 이러한 예측지표는 고위험군의 환자에게는 항암치료를 곁들인 좀더 적극적인 치료를, 저위험군의 환자에게는 쓸데없는 과도한 치료를 막는 중요한 역할을 한다.

수만개 유전자들의 cDNA 마이크로어레이 발현 자료 뿐 아니라 축적되어 있는 그들의 생물학적 자원들을 이용하여 환자의 예후나 생존을 예측할 수 있는 지표 개발을 위해서 다음과 같은 과정들을 고려하여야 한다.

유전자군의 조성

종양을 치료할 수 있는 신약 개발은 우선 종양 발생과 관련된 유전자를 찾고 이 유전자의 기능을 규명한 후 여기에 영향을 미치는 약물을 개발하는 순서로 진행된다. 그러나 종양의 종류에 따라서는 정상 세포가 종양 세포로 전이되는 과정이 꽤 복잡하기 때문에 유전자를 찾기 보다는 변화가 일어나는 대사경로를 찾는 것도 중요하다. 모든 생물의 세포 내에서 정확히 조절되고 있는 대사경로에 돌연변이가 발생하면 세포가 발달하는 방식에 변화가 생기고 그 결과로 종양이 발생하기 때문이다. 그러므로 종양 발생에 영향을 미치는 대사경로를 찾을 수 있다면 암세포 사멸을 유도할 수 있고 또 새로운 유형의 종양 치료법 개발이 가능해 진다.

생물학 전문가들의 작업에 의해 유전자들의 기능이나 대사경로에 관련된 정보가 구축된 KEGG(Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp./kegg), GO(Gene Ontology, http://www.geneontoloty.org)와 MBGD (Microbial Genome Database for Comparative Analysis, http://mbgd.genome.ad.jp) 등의 데이터베이스를 이용하면 각 유전자들이 어느 대사경로에 속하여 있는지 알아낼 수 있고 이들 대사경로를 단위로 유전자 집합 분석을 진행할 수 있다.

개별유전자의 특이발현성 검정

개별유전자를 대상으로 표현형과 유의한 관련이 있는 특이발현유전자를 찾는 연구는 활발히 진행되어 많은 결과들이 나와 있다. 제일 단순한 형태의 fold change rule을 시작으로 t-통계량을 이용한 방법, Tusher 등 (2001)이 제안한 SAM(Significance Analysis of Microarrays), Tibshirani 등 (2003)의 PAM(Prediction Analysis for Microarrays)과 principal component analysis, Kerr과 Churchill (2001)의 ANOVA 모형을 사용하는 방법 등이 있다. 그러나 생존자료인 경우에는 단변량 Cox 모형을 이용한 Wald 검정이 유전자가 생존시간에 영향을 끼치는지 판단할 수 있는 제일 좋은 방법이다.

대사경로의 유의성 검정

마이크로어레이 자료에서 '어떤 기능, 또는 어떤 대사경로를 수행하는 유전자 집합이 가장 중요한가?'라

는 생각이 들 수 있다. 같은 대사기능을 수행하는 유전자들을 함께 묶어 이들이 질병의 발생이나 수명에 영향을 미치는지 밝히는 대사경로 분석은 카이제곱 검정, 오즈비와 초기하분포를 이용한 z-score 검정 등으로부터 분석대상 대사경로에 속한 유전자들과 특이발현유전자군 사이의 관련 여부를 따질 수 있다. 그러나 이것은 특이발현유전자군의 원소가 관심 유전자군에 포함된 여부만을 따지는 것이고 여기서 한걸음 더 나아가 상위 특이발현유전자가 관심 유전자군에 어떻게 포함되었는지 분석에 반영하는 GSEA(Gene Set Enrichment Analysis) (Mootha 등, 2003; Subramanian 등, 2005)가 개발되어 많이 쓰이고 있다. 또한 score 검정을 이용하여 관심 유전자군이 환자들의 이진표현형과 관련있는지 검정하는 global test가 유도되었고 (Goerman 등, 2004), 이 아이디어를 각 대사경로가 환자들의 생존율과 관계있는지 검정가능한 통계량으로 영역을 확대하였다 (Goerman 등, 2005). 그리고 중심극한정리를 이용한 모수적 방법인 PAGE(Parametric Analysis of Gene set Enrichment) (Kim과 Volsky, 2005)와 SAM의 t-통계량을 유전자군 분석으로 확장한 SAM-GS (Dinu 등, 2007)등의 대표적 방법이 있으나 이들은 생존자료 분석에는 사용할 수 없는 단점이 있다. Adewale 등 (2008)은 유전자들과 표현형 간의 회귀모형을 바탕으로 Wald 통계량을 이용한 검정법을 제시하였는데 생존분석의 경우 Cox 모형을 이용한 적용이 가능하다.

대표 주성분을 이용한 생존예측 지표개발

하나의 대사경로는 적게는 두세 개, 많게는 천 개 가까운 유전자들로 구성되고 biological process의 범주에서 찾을 수 있는 대사경로의 수만 하여도 4000개가 넘는 것으로 알려져 있다. 표본의 수에 비해 유전자수가 월등히 많은 마이크로어레이 자료의 특성상 여러 대사경로들을 이용하여 지표개발을 하기 위해서는 대사경로에 속한 각 유전자들의 정보를 모두 독립적으로 표현하는 것은 불가능하고 이들을 축소하여 대표값으로 표현하는 방법을 찾아야 한다. 보통 관찰치 집단에서 하나의 대표값을 추리는 방법으로 모든 관찰값들의 평균, 중간값 등이 많이 사용되고 있으나, 지표개발의 목적 아래에서는 적당하지 않다. 또한 대사경로 내에서 특이발현 정도가 제일 큰 유전자의 발현값을 대표값으로 할 수도 있지만 정보의 손실을 가능한 작게 하면서 축약하는 또 하나의 방법으로 주성분분석(principal component analysis; PCA)이 있다. 이 분석을 통하여 상관이 있는 유전자들을 적절히 선형결합하여 의미있고 서로 독립적인 소수의 주성분으로 변환시키는 것이 가능하며 대표 주성분값을 각 대사경로의 대표값으로 사용할 수 있다. 대사경로에 속한 모든 유전자를 사용하여 PCA를 하는 대신 Chen 등 (2008)은 표현형과 강한 연관이 있는 유전자들만을 대상으로하는 supervised PCA 방법을 제시하였다.

대사경로의 대표값 설정

중도절단이 포함된 생존자료의 회귀모형으로 가장 많이 쓰이는 것은 당연히 Cox의 비례위험모형이다. 환자에수보다 훨씬 많은 대사경로의 대표값들을 동시에 이용하여 모형을 적합할 때 생길 수 있는 다중공선성문제 해결을 위하여 벌점(penalty)을 이용하여 상대적으로 유의하지 않은 대사경로의 영향을 줄여나가는 regularization 방법이 개발되었다. L_2 벌점을 이용한 능형회귀, L_1 벌점을 이용하여 변수선택을 하는 Tibshirani (1997)가 제안한 Lasso와 cluster threshold gradient descent regularization 등의 방법이 있다. 이외에도 많은 수의 대표값을 대상으로 전진선택, 후진제거나 단계선택 등의 방법으로 변수를 결정하고 모형을 추정할 수 있다.

지표의 정확성 평가

생성된 진단모형이나 지표의 예측정확성을 바르게 평가하기 위해 표본 일부를 훈련군으로 설정하여 모형을 만들고 나머지 자료를 시험군으로 하여 정확성을 검증하는 2단계 교차검증 작업을 수행하는 것이

바람직하다. 주어진 표본들을 서로 독립적인 훈련군과 시험군으로 나누는 것이 가장 바람직하지만 표본의 수가 많지 않은 마이크로어레이 자료 특성상 k-fold 교차검증과 leave one out 교차검증을 사용할 수 있다. Simon 등 (2003)은 변수 선택, 모형 설정과 검증의 과정이 제대로 조절되지 않고 시험군의 정보가 검증 절차 이전에 사용될 경우 과적합(overfitting)으로 인한 오류가 발생할 수 있다는 것을 보였다.

생성된 예측지표의 정확성 평가를 위한 몇 가지 척도가 있다. 가장 보편적인 방법은 지표를 이용하여 시험군의 환자들을 고위험군과 저위험군으로 분류하고 두 군 사이에 유의한 생존율 차이가 있는지 로그순위검정을 하는 것이다. Heagerty 등 (2000)은 모형의 정분류율과 오분류율의 변화를 나타내는 ROC(receiver operating characteristic) curve를 시간에 따라 변화하는 환자들의 상태를 표현할 수 있는 time-dependent ROC curve로 확장하였다. Graf 등 (1999)는 이진표현형의 실제 결과와 예측 생존확률의 차이에 대한 평균제곱편차를 이용한 Brier Score (Brier, 1950)를 중도절단이 존재하는 생존자료에 대한 지표의 정확성을 평가할 수 있도록 확장하였다.

3. DLBCL 환자들의 생존예측지표개발

본 논문은 240예의 DLBCL 환자들을 장기간 관찰한 생존시간, 마이크로어레이 자료와 대사경로 정보를 이용하여 환자들의 생존예측지표를 찾는 pilot study 결과를 보인다.

1) DLBCL 자료

웹상에 공개된 마이크로어레이 자료는 꽤 있지만 표현형이 생존시간인 자료는 많지 않다. 그중에서 환자예수가 많고 Rosenwald 등 (2002)와 Bair과 Tibshirani (2004)의 논문에서 인용된 DLBCL 자료를 분석에 사용하였다.

DLBCL 자료는 처음으로 DLBCL 판정을 받은 환자 240명으로부터 채취한 생검자료와 그들이 항암치료를 받고 사망까지의 시간을 관찰한 자료를 일컫는다. 생검자료로부터 마이크로어레이실험을 통하여 모두 7399개의 발현값을 관찰하였지만 일부는 전처리과정에서 제거되었고 중복된 유전자들은 median 값을 사용하도록 정리하니 4443개로 축소되었다. 이들을 대상으로 KEGG를 이용하여 1389개 유전자들과 관련된 193개의 대사경로 정보를 모았다.

240명 환자들의 생존시간은 0∼21.8년, 관찰 추적시간의 중앙값은 2.8년(생존자는 7.3년)이며 관찰기간 중에 138명(57%)이 사망하였다. 또한 임상변수에 의하면 병기가 1기인 환자가 15%, 2기 31%, 3기 20%, 4기 34%로 분류되었다.

2) 기본적인 방법 제안과 모형 I의 분석 결과

DLBCL 자료는 KEGG를 이용하여 대사경로에 따른 유전자군을 조성하였고, 각 대사경로로부터 대표 주성분을 뽑아낸 다음 그들 각각의 유의성을 검정하는 방법으로 2절에서 나열한 대사경로의 유의성 검정과 대표값 설정의 단계를 대신하였다.

다음의 기본적인 2단계 과정을 통하여 생존예측지표를 찾아보았다.

(i) 각 대사경로의 차원축소와 대표 주성분의 유의성 검정

전체 유전자수에 비해 각 대사경로에 속한 유전자의 수는 훨씬 작아졌지만 아직도 small 'n', large 'p'의 문제는 여전하기 때문에 주성분분석을 통한 차원축소를 실시하였다.

대사경로에서 생성된 주성분들의 상대적 중요 순위는 각 주성분이 원래 유전자가 갖고 있는 정보를 설명하는 정도에 따라 정해지므로 생존시간과 관련여부를 나타내는 순위와 일치하지는 않는다.

표 **3.1.** 모형 I과 II를 사용한 생존예측지표 추정 결과

	모형 I	사전 유전자 걸러내기 실시	
		모형 II-1	모형 II-2
분석대상 유전자수	1389	238	885
분석대상 대사경로 수	193	141	168
유의한 대표 주성분의 수	12	60	21
지표에 포함된 대표 주성분의 수	5	6	7
지표에 포함된 유전자의 수	82	55	131
로그 순위검정의 p-value	0.045	0.0452	0.605

각 대사경로의 대표값으로 어떤 주성분을 몇 개까지 채택할 것인가 하는 기준이 필요한데 본 분석에서는 단변량 Cox 모형을 이용하여 생존시간과 유의한 관련이 있는 주성분($p < 0.05$)들 중 제일 영향을 크게 미치는 주성분을 대표로 선택하였다. 이러한 방법은 주성분회귀분석 (Hastie 등, 2001, Chapter 3.4.4)에서 사용되는 주성분들이 실제로는 생존시간과 무관한 단점을 개선하는 의미가 있다.

(ii) 생존예측지표 개발과 정확성 검증

생존분석에서 가장 대표적인 통계적 모형은 Cox의 비례위험모형이다. 개별유전자 분석을 이용한 지표개발에서는 regularization을 이용한 모형추정이 많이 사용되지만 대표 주성분들의 수는 대사경로의 수 보다 적어 고차원의 부담도 줄고 또 지표 구성에 사용된 유전자들에 대한 해석을 간단히 하기 위하여 단계적으로 변수를 선택(stepwise selection)하는 방법을 이용하여 생존예측지표를 개발하였다.

개발된 지표의 성능 확인을 위하여 교차검증(cross validation)이 필수적이다. 240예 중 임의로 선택한 160예의 훈련군 자료를 이용하여 대표 주성분 결정과 생존예측지표를 개발하였고 시험군에 속한 80예가 고위험군과 저위험군으로 제대로 분리되었는지 로그순위검정을 통하여 예측의 정확성을 확인하였다.

위의 두 단계를 거쳐서 만들어진 예측지표를 모형 I이라 하자. 모형 I은 훈련군 환자들의 유전자 발현과 종양진단후 추적한 생존시간의 정보로부터 12개 대사경로의 유의한 대표 주성분들을 찾아냈고 이들 중 5개의 주성분값이 단계선택방법으로 변수채택되어 생존예측지표를 구성하였다(표 3.1). 이 지표로부터 시험군 환자들의 risk score를 계산하고 중앙값을 기준으로 고위험군과 저위험군으로 나눈 후 로그순위검정을 실시한 결과, 두 군의 생존율 사이에 유의한 차이가 있는 것으로 나타났다(그림 3.1, p-value = 0.045)

3) 사전 유전자 걸러내기와 모형 II의 분석 결과

마이크로어레이 시험에서 전체 유전자수는 많지만 실질적으로 표현형에 유의한 영향을 주는 유전자는 그리 많지 않다. 전체 유전자를 대상으로 모형 I을 구축하기 전에 유전자 걸러내기(gene filtering)를 실시하는 다음의 두 가지 형태의 모형 II를 제안한다.

모형 II-1. 생존시간에 영향을 주는 유전자만 선택 후 모형 I 추정.

생존시간의 차이에 유의한 영향을 주는 유전자들만 대상으로 지표를 추정하기 위하여 단변량 Cox모형을 적용하여 환자의 생존시간에 유의한 영향(p-value < 0.05)을 주는 238개 유전자를 선택한 후 이들을 대상으로 생존예측지표를 추정하였고(표 3.1) 환자들을 고위험군과 저위험군으로 잘 나눔을 확인하

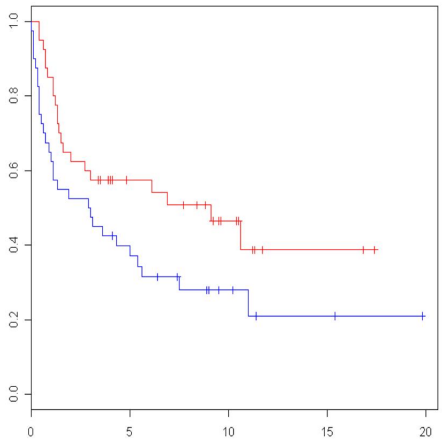
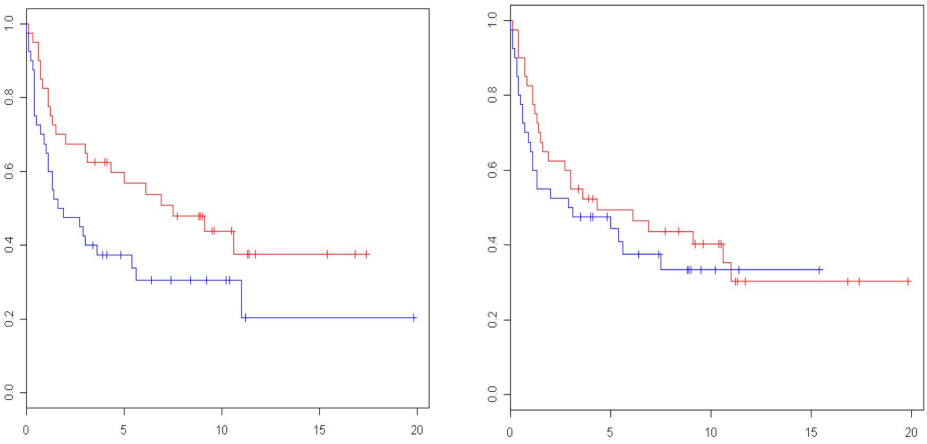


그림 3.1. 모형 I을 사용하여 시험군 80례를 고위험군과 저위험군으로 나눈 Kaplan-Meier 생존곡선



(a) 모형 II-1 (b) 모형 II-2

그림 3.2. 모형 II를 사용하여 시험군 80례를 고위험군과 저위험군으로 나눈 Kaplan-Meier 생존곡선

였다(그림 3.2(a) p-value $= 0.0452$).

전체 유전자를 대상으로 193개 대사경로로부터 유의한 대표 주성분을 찾을 때는 12개밖에 없었지만 생존시간에 유의한 영향을 주는 유전자들을 대상으로는 60개를 찾을 수 있었다. 그런데 전체 유전자를 대상으로 찾은 12개의 대표 주성분을 배출한 대사경로중 6개가 후자와 겹치는 것으로 유전자 걸러내기의 영향이 매우 큰 것을 알 수 있었다.

모형 I과 모형 II-1의 최종지표에 포함된 대표 주성분이 나타내는 대사경로들을 비교하면 동일한 대사경로는 없었다. 그러나 KEGG 검색을 해보면 대사경로들 사이에 상당한 포함관계가 존재하는 것을 볼 수 있었고 모형 I과 모형 II-1의 최종지표에 각각 포함된 82개와 55개 유전자중 17개가 겹치는 것을 발견하였다.

이 외에도 특이발현 유전자를 선택하는 방법으로 중도절단된 환자들을 제외한 환자들의 생존시간의 중앙값을 기준으로 두 군으로 나누어 SAM의 t 통계량을 이용하여 186개를 선택하였다. 이들을 대상으로 모형 I의 지표를 찾았지만 좋은 결과를 얻지 못하였다(p-value $= 0.494$).

모형 II-2. 생존시간에 영향을 미치지 않는 유전자를 제거 후 모형 I 추정.

개별유전자 분석에서는 잡음의 효과를 보일 수 있는 비특이발현 유전자를 전처리과정에서 미리 제외시킴으로서 분석 오류 요인을 제거하고 동시에 분석대상 유전자 수를 줄이는 효과를 얻기도 한다. 1399개 유전자 중 1/3에 해당하는 변동계수(coefficient of variation)가 10 이하인 유전자를 제거하고 885개의 유전자를 분석하였다(표 3.1). 그러나 여기서 추정된 생존예측지표는 환자들을 고위험군과 저위험군으로 나누는데 실패하였다(그림 3.2(b), p-value $= 0.404$). 대사경로 별로 분석을 하는 경우에는 분석 대상 유전자 수가 크지 않아서 사전 유전자 걸러내기가 효과를 보지 못한 것이라 생각한다.

4. 토론

마이크로어레이 자료로부터 지표를 개발하기 위한 최적의 분석 방법은 질병의 종류와 표현형의 형태에 따라 차이가 있다. 본 논문에서 다룬 DLBCL 자료 분석은 최적의 방법을 적용하였다고 할 수 없고 추정된 생존예측지표는 성능이 뛰어나지도 않지만 장기간 추적한 생존자료와 그들의 유전자 정보 뿐만 아니라 데이터베이스에 축적된 생물학적 정보를 이용한 대사경로 차원에서 분석하였다는 점에서 충분히 의미가 있다.

전처리과정을 거친 후 약 30% 정도의 유전자만 KEGG에서 관련 정보를 찾아 볼 수 있었고 나머지 유전자들은 분석에서 제외되었다. 대사경로 정보를 갖지 않는 유전자들에 대하여 k-평균 군집화 기법을 적용하여 k개의 집합으로 분류하여 대사경로 분석에 추가하는 방법도 거론되고 있지만 실제 자료분석에서 의미있는 결과를 얻었다는 보고는 없었다 (Chen과 Wang, 2009). KEGG에 정보가 들어있지 않는 유전자들은 대부분 특별한 기능이 없는 유전자들이기 때문일 것이다.

다수의 대사경로들 사이에는 서로 포함관계를 이루는 경우도 있고 강한 상관관계가 존재하기도 한다. 그런데 대상 대사경로들로부터 생존과 유의한 관련이 있는 대표값을 만들어내는 방법은 단변량적 접근이므로 대사경로 사이의 관계는 전혀 고려되지 않았다. 그러므로 이러한 점을 개선할 수 있도록 상관관계가 높거나 중복되는 정보를 갖는 대사경로들은 지표를 만들기 전 미리 정리되는 것이 바람직하겠다. 또한 유전자 걸러내기가 전체 결과에 큰 변화를 주는 것을 볼 수 있는데 이에 대한 구체적인 효과 분석에 대한 심도있는 연구가 필요하다.

참고문헌

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

Bair, E. and Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene Downloaded from gene expression data, *PLOS Biology*, **2**, 511–522.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability, *Monthly Weather Review*, **78**, 1–3.

Chen, X. and Wang, L. (2009). Integrating Biological Knowledge with Gene Expression Profiles for Survival Prediction of Cancer, *Journal of Computational Biology*, **16**, 265–278.

Chen, X., Wang, L., Smith, J. D. and Zhang, B. (2008). Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes, *Bioinformatics*, **24**, 2479–2481.

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.

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.

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.

Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data, *Statistics in Medicine*, **18**, 2529–2545.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2001). *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, Springer-Verlag, New York.

Heagerty, P. J., Lumley, T. and Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker, *Biometrics*, **56**, 337–344.

Kerr M. and Churchill, G. (2001). Experimental design for gene expression microarrays, *Biostatistics*, **2**, 183–201.

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

Ma, X. J., Wang, Z., Ryan, P. D., Isakoff, S. J., Barmettler, A., Fuller, A., Muir, B., Mohapatra, G., Salunga, R., Tuggle, J. T., Tran, Y., Tran, D., Tassin, A., Amon, P., Wang, W., Wang, W., Enright, E., Stecker, K., Estepa-Sabal, E., Smith, B., Younger, J., Balis, U., Michaelson, J., Bhan, A., Habin, K., Baer, T. M., Brugge, J., Haber, D. A., Erlander, M. G. and Sgroi, D. C. (2004). A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen, *Cancer Cell*, **5**, 607–616.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D. and Groop, L. C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes, *Nature Genetics*, **34**, 267–273.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Montserrat, E., Lopez-Guillermo, A., Grogan, T. M., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T. and Staudt, L. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma, *The New England Journal of Medicine*, **346**, 1937–1947.

Simon, R., Radmacher, M. D., Dobbin, K. and McShane, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification, *Journal of National Cancer Institutes*, **95**, 14–18.

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. and Mesirov, J. P. (2005). Gene set enrichment analysis:A knowledge-based approach for interpreting genome-wide expression profiles, *PNAS*, **102**, 15545–15550.

Tibshirani, R. (1997). The Lasso method for variable selection in the cox model, *Statistics in Medicine*, **16**, 385–395.

Tibshirani, R., Hastie, T. Narasimhan, B. and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays, *Statistical Science*, **18**, 104–117.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response, *PNAS*, **98**, 5116–5121.

Predicting Survival of DLBCL Patients in Pathway-Based Microarray Analysis

Kwanghyun Lee[1] · Sunho Lee[2]

[1]Department of Applied Statistics, Sejong University
[2]Department of Applied Statistics, Sejong University

(Received April 2010; accepted May 2010)

Abstract

Predicting survival from microarray data is not easy due to the problem of high dimensionality of data and the existence of censored observations. Also the limitation of individual gene analysis causes the shift of focus to the level of gene sets with functionally related genes. For developing a survival prediction model based on pathway information, the methods for selecting a supergene using principal component analysis and testing its significance for each pathway are discussed. Besides, the performance of gene filtering is compared.

Keywords: Microarray experiment, survival analysis, pathway, principal component analysis, proportional hazards model.

This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD, Basic Research Promotion Fund)(KRF-2008-531-C00019).

[2]Corresponding author: Professor, Department of Applied Statistics, Sejong University, Gunjadong, Kwangjingu, Seoul 143-747, Korea. E-mail: leesh@sejong.ac.kr q