The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.2.307
26(2), 307–319

Cross Platform Data Analysis in Microarray Experiment

Jangmee Lee[a] · Sunho Lee[a,1]

[a]Division of Mathematics and Statistics, Sejong University

(Received January 7, 2013; Revised March 26, 2013; Accepted March 26, 2013)

Abstract

With the rapid accumulation of microarray data, it is a significant challenge to integrate available data sets addressing the same biological questions that can provide more samples and better experimental results. Sometimes, different microarray platforms make it difficult to effectively integrate data from several studies and there is no consensus on which method is the best to produce a single and unified data set. Methods using median rank score, quantile discretization and standardization (which directly combine rescaled gene expression values) and meta-analysis (which combine the results of individual studies at the interpretative level) are reviewed. Real data examples downloaded from GEO are used to compare the performance of these methods and to evaluate if the combined data set detects more reliable information from the separated data sets or not.

Keywords: Microarray experiment, cross-platform, integration, median rank score, quantile discretization, standardization, meta analysis.

1. 서론

마이크로어레이 기술은 1995년 미국 Stanford 대학교에서 개발되어 종양에 대한 유전학적 특성과 기전 연구를 활성화했고 질병의 진단과 치료를 하는데 크게 이바지하였다. 동시에 수만 개 유전자를 관찰할 수 있는 마이크로어레이 실험 자료의 가치와 공유의 필요성이 주목받으면서 1999년에 미국 National Center for Biotechnology Information의 Gene Expression Omnibus(GEO, http://www.ncbi.nlm.nih.gov/geo), Stanford 대학교의 Stanford Microarray Database(SMD, http://smd.stanford.edu), 2002년 European Bioinformatics Institute의 ArrayExpress(http://www.ebi.ac.uk/arrayexpress) 등이 구축되어 공개 데이터 저장소의 역할 뿐만 아니라 데이터 검색과 해석을 위한 도구도 지원하고 있다. 또한, 유명 저널들에서 각 논문에 인용된 데이터를 공개 저장소에 올리는 것을 관행화하면서 현재 GEO에는 전 세계 5,000여 연구소로부터 제공받은 10,000 종류가 넘는 플랫폼으로 구성된 700,000개 이상의 표본을 대상으로 수행한 30,000개가 넘는 연구 정보가 등록되어 있고 하루에 약 60,000건의 데이터 검색이 이루어지고 있다. 그런데 마이크로어레이 실험의 조건이 까다롭고 큰 비용이 들어 보통 소규모의

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(No. 2012-0004330).
[1]Corresponding author: Professor, Division of Mathematics and Statistics, Sejong University, 98, Gun-jadong, Kwangjinku, Seoul 143-747, Korea. E-mail : leesh@sejong.ac.kr

실험이 시행되므로 표본 수가 50개 이상인 연구는 11.58%, 30개 이상은 21%에 불과하다. 표본이 많지 않으면 강건한(robust) 정보를 얻을 수 없을 뿐만 아니라 바이어스로 인한 치우친 결과를 얻기 쉽고 대표본에서의 중심극한 정리를 가정할 수 없는 아쉬움이 있다.

서로 독립적으로 진행되었지만 실험목적이 같은 여러 연구 자료를 저장소에서 다운받아 한 개의 커다란 자료집합으로 통합하여 분석한다면 표본수의 증가로 검정력이 높아지며, 단일 연구에서 예상치 않게 발생할 수 있는 오류가 보완될 수 있으며 여러 연구를 합한 결과는 일반화하기 쉽고 파급효과도 커질 것이다. 그러나 실험이 진행된 환경과 조건, 마이크로어레이칩에 점적된 유전자의 종류가 동일하지 않은 데이터들을 통합하는 솔루션이 완전하지 않으며, 특히 마이크로어레이 플랫폼이 서로 다른 경우 데이터 내부 구조의 차이로 이를 통합하여 분석하는데 많은 문제점이 있다.

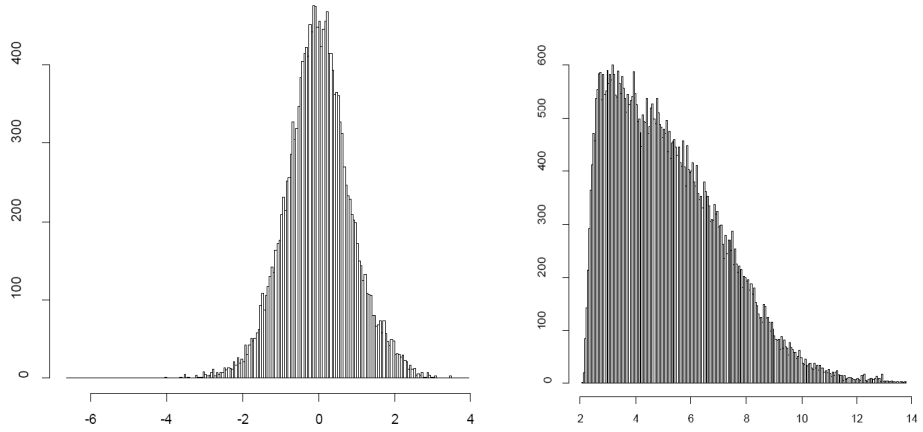
이미 플랫폼이 다른 연구들을 병합하여 분석하는 여러 가지 방법이 제시되었으나 상대적인 비교 연구는 충분치 않았기 때문에 본 논문에서는 이진 표현형을 갖는 자료를 대상으로 이들의 예측 정확성을 사용하여 통합의 필요성과 통합 방법의 우수성을 비교해 보려 한다. 2절에서는 서로 다른 연구 자료를 통합할 때 생기는 문제점에 대하여 설명하고 3절에서는 현재 사용되고 있는 통합 방법을 알아본다. 4절에서는 공용 데이터 저장소에서 다운받은 실제 자료로부터 구축한 분류자의 예측 정확성을 이용하여 각 통합 방법의 효율성의 차이를 알아본다.

2. 서로 다른 연구 자료 통합에 따른 문제점

연구에 따라 제공되는 마이크로어레이 실험 자료의 포맷이 모두 같지 않아서 서로 다른 데이터의 상호 접근과 공유가 쉽지 않다. 이런 문제점을 해결하기 위한 노력으로 Microarray Gene Expression Data라는 단체에서는 마이크로어레이 실험 관련 자료에서 기록해야 할 최소한의 필수적인 내용을 정의한 Minimum Information About a Microarray Experiment(MIAME)의 기준을 제시 (Brazma 등, 2001)하여 실험 자료의 표현과 공유를 위한 표준화 작업을 진행하였고 여러 저널과 데이터 저장소에서도 MIAME 기준을 따르도록 권고하고 있다.

데이터 포맷의 차이 외에도 자료 공유에는 여러 난관이 있다. 사용된 마이크로어레이칩이 다를 경우 각 칩에 점적된 유전자들 목록도 다르고 유전자 ID를 표현하는 방법도 Clone ID, Probe ID, Gene name, GenBank Accession이나 Unigene Cluster ID 등 다양하기 때문이다. 이들 유전자 ID간에 상호매핑(cross mapping)은 필수이나 자동으로 이루어지지 않고 있고 SOURCE(http://source.stanford.edu; Diehn 등, 2003), BioGPS(http://biogps.gnf.org; Wu 등, 2009)와 DAVID(http://david.abcc.ncifcrf.gov; Huang 등, 2009) 등을 이용하여 공통적인 ID로 매핑한 후 통합하려는 연구에 모두 속한 유전자들을 대상으로 분석을 진행한다. 이런 매핑과정에서 서로 다른 형태의 ID간에 관계가 규명되지 않은 유전자도 있지만 중복으로 규명된 경우도 많아 정보의 손실이 발생한다. 공통 ID와 관계 규명이 안 된 유전자들은 제거되고, 공통 ID에 기존의 ID들이 중복으로 표현된 경우에는 중복 유전자들의 평균이나 중앙값을 이용하여 한 개의 값으로 축소되기 때문이다.

서로 다른 연구에서 얻어진 표본들을 합치는데 제일 큰 어려움은 마이크로어레이 플랫폼이 다르면 관찰값의 분포 형태도 다르다는 문제이다. 대표적인 플랫폼으로는 마이크로어레이칩을 만드는 데 사용하는 검출용 염기 서열에 따라 cDNA(200–500 염기쌍)칩과 올리고 염기서열(15–100 염기쌍)칩을 들 수 있는데 이들 모두 유전자의 발현 강도를 측정하지만 어레이를 만든 원리와 강도를 재는 척도가 다르다. Figure 2.1은 GEO에서 다운받은 대장암 관련 cDNA 자료(GSE20970, 종양 43예와 정상 30예)와 올리고 자료(GSE23878, 종양 35예와 정상 24예)에 각각 속한 유전자들의 평균값 분포를 나타낸 히스토그램이다. cDNA 칩은 준거자료(reference sample)와 비교한 발현비에 로그 변환을 하여 0을 중심으로 대



(a) cDNA micro array (b) Oligonucleotide array

Figure 2.1. Histogram of mean expression value of each gene in different platforms

칭인 분포를 보이지만, 올리고칩은 11개 내지 20개의 probe들의 대푯값으로 보통 5 이상의 실수값을 가지며 서로 분포의 형태도 다름을 볼 수 있다.

서로 다른 플랫폼에 관한 비교 연구가 많이 진행되었는데 플랫폼에 상관없이 유전자들의 발현 형태는 서로 일치한다는 결과 (Larkin 등, 2005; Shi 등, 2006)와 그렇지 않다 (Kuo 등, 2002; Tan 등, 2003)는 상반된 결과가 있다. Stec 등 (2005)은 동일한 표본에 대하여 cDNA칩과 올리고칩을 각각 사용하여 실험하였을 때 공통 유전자들의 발현값 사이에 상관관계가 낮을 뿐 아니라 음의 상관관계를 갖는 유전자도 30% 가까이 된다는 것을 보였다.

3. 서로 다른 연구 자료의 병합

하나로 통합하려는 대상 연구들이 서로 다른 환경과 조건에서 실험이 시행되었을 경우 공통 유전자들의 관찰값 분포가 서로 다르므로 각 연구에 속한 표본들의 자료를 아무런 과정을 거치지 않고 그냥 통합하는 것은 불가능하다. 통합 대상의 모든 표본의 관찰값들이 같은 분포를 하도록 변환하는 방법으로 순위에 따른 중위수(median rank score; Warnat 등, 2005) 변환, 분위수를 이용한 이산화(quantile discretization; Liu 등, 2002)와 표준화(standardizaton) 방법을 사용할 수 있고 원자료값을 변환하는 대신 분석 결과를 병합하는 메타분석법 (Fisher, 1932)이 있다.

순위기반 중위수(median rank score; MRS) 변환

Warnat 등 (2005)이 사용한 방법으로 순위를 기반으로 하여 서로 다른 형태 자료들의 관찰값 분포가 같도록 변환시킨다. 통합 대상의 연구 중 표본이 가장 많은 연구를 준거집합(reference set)으로 지정한 후 그 안에 속한 표본을 대상으로 각 유전자의 중위수를 구하고 순위를 매긴다. 비준거집합(non-reference set)에 속한 각 표본의 분석대상 유전자들의 발현값을 표본 내에서 순위를 정한 후 그 순위에 대응하는 준거집합의 중위수로 자료값을 대치한다. 그러므로 이상점의 영향이 감소하는 경향이 있다.

분위수를 이용한 이산화(quantile discretization; QD)

자료의 이산화(discretization)는 대용량 자료의 데이터마이닝을 위한 전처리 과정으로 자료의 구조를

쉽게 이해하고 분석 수행 속도를 높이기 위하여 많이 사용된다. Liu 등 (2002)은 연속형 관찰값을 크기 순으로 나열하여 인접한 구간을 분할하거나 통합하는 방법에 관한 연구를 하였으며 이 중에서 제일 간단한 방법으로 대상 자료들을 크기대로 정렬한 후 자료 수를 중심으로 k 등분하고 같은 구간에 속한 자료들에 동일한 값을 주는 이산화 방법을 제시하였다. Warnat 등 (2005)은 서로 다른 플랫폼 자료의 통합에 이산화 방법을 도입하여 각 표본에 속한 유전자들을 8개의 군으로 등분한 후 각 군에 속한 관찰값들을 각각 $-3, -2, -1, 0, 0, 1, 2, 3$으로 대치하는 분위수 이산화(quantile discretization)를 선보였고 이러한 이산화는 플랫폼의 차이에서 비롯되는 변동성의 차이도 함께 단순화하기 때문에 바람직하다고 하였다. Kim 등 (2008)은 각 유전자의 순위를 이용한 이산화를 시도하였고 이러한 이산화는 정보의 손실은 있지만 특이점에 로버스트하며 계산이 간단하고 이해하기 쉬운 장점이 있다고 하였다.

표준화(standardizaton; STD)

각 표본에 속한 유전자들의 발현값 평균이 0, 분산이 1이 되도록 표준화하는 방식, 즉 z-score를 이용하여 분석 대상의 모든 표본의 관찰값 분포를 변환시키는 방법으로 매우 간단하고 유전자간 순위정보도 유지하며 어레이간의 scale 보정이 필요 없다는 장점이 있다.

위에 언급한 방법들은 각 실험 자료의 플랫폼이나 정규화(normalization) 방법과 상관없이 표본들을 표현형별로 직접 통합이 가능하다. 중위수 변환법이나 표준화를 이용한 변환은 이산화보다 유전자들 사이의 순위 상관관계도 유지하고 정보의 손실이 덜한 장점이 있다. 그렇지만 Warnat 등 (2005)은 분위수 이산화를 이용한 자료의 축약은 정보를 손실시키기만 하는 것이 아니고 연구 환경이나 플랫폼의 차이에 따른 변이도 함께 감소시키며, 몇 가지 자료 분석 예를 통하여 이산화 방법이 중위수 변환을 이용한 통합보다 표현형에 대한 예측 정확도가 높다는 것을 보였다. 위에 열거한 방법 외에도 서로 다른 연구에 따른 관찰값의 분포의 차이를 선형모형으로 설명한 cross platform normalization(XPN; Shabalin 등, 2008)과 Empirical Bayes method (Walker 등, 2008)도 있다. Rudy와 Valafar (2011)는 실제 자료 집합을 이용한 분석에서 처리군의 크기가 거의 같은 경우 XPN 방법이 서로 다른 플랫폼의 차이를 잘 설명하는 것을 보였다.

메타분석(meta analysis; META)

여러 연구 결과를 종합하여 하나의 결론을 얻는 메타분석은 각 연구에서 얻은 원자료들을 통합하는 대신 유의확률이나 효과크기(effect size)를 하나로 통합하는 방법이다. 마이크로어레이 자료들의 메타분석 (Hong과 Breitling, 2008; Campain과 Yang, 2010)은 유전자들 사이의 발현값과 순위 정보를 잃게 되고 자료 집합들의 질에 크게 영향을 받는다는 단점이 있지만 통합하려는 플랫폼이 달라서 생기는 문제가 자연스레 해결되는 장점이 있다.

메타분석에도 여러 가지가 있지만 제일 기초적인 방법은 Fisher (1932)가 제안한 역카이제곱법(inverse chi-square method)으로 i번째 연구($i = 1, 2, \ldots, k$)에서 특정 유전자의 특이발현 여부에 대한 유의확률이 p_i였다면 k개 연구의 결합 결과 $-2\sum \log p_i$는 자유도가 $2k$인 카이제곱분포를 따른다는 것이다. Good (1955)은 각 연구에 참여한 표본의 수나 유의확률에 대한 신뢰도를 가중치로 사용하여 Fisher의 방법을 확장하였고 Campain과 Yang (2010)은 8가지 메타분석법의 비교에서 서로 플랫폼이 다른 경우에는 Fisher의 방법을 사용하였을 때 예측 정분류율이 제일 높음을 보였다.

4. 실제 자료 분석

GEO 검색을 통하여 Diffuse Large B-Cell Lymphoma(DLBCL)와 관련된 3개의 연구 결과와 폐암에 관한 3개의 연구 결과를 찾았다. 두 가지 종양의 실제 자료 분석을 통하여 서로 다른 플랫폼으로 구성된

Table 4.1. Details of DLBCL microarray studies

Name	Study	Platform	Samples	Probes	GEO
[C1]	Alizadeh *et al.* (2000)	cDNA	14 GCB, 13 ABC	18432	GSE60
[O1]	Williams *et al.* (2010)	Oligo	27 GCB, 21 ABC	54675	GSE19246
[O2]	Shaknovich *et al.* (2010)	Oligo	40 GCB, 20 ABC	54675	GSE23501

연구들의 효율적인 변환과 통합 방법, 그리고 통합의 필요성을 알아보았다.

3절에서 언급한 네 가지 방법에 따라 훈련군의 정보를 통합하였을 때 시험군에 속한 표본의 표현형 예측에 어떤 차이가 있는지 알아보기 위하여 DLBCL과 폐암 자료 각각에 대하여 모든 유전자 발현값을 중위수 변환, 이산화 변환과 표준화를 이용하여 각각 변환한 후, 한 개 연구는 시험군으로, 나머지 두 연구는 통합하여 훈련군으로 사용하였다. 메타 분석은 사전에 훈련군과 시험군의 모든 자료를 표준화시켜 사용하였다. 훈련군에서 구한 특이발현 유전자를 이용하여 시험군에 속한 표본들의 표현형을 예측하였고 이들 정분류율(classification accuracy)을 근거로 통합 방법에 따른 효율성을 판단하였다. 또한, 플랫폼이 다른 연구들의 통합이 정분류율을 높이는데 실질적으로 도움이 되는지 알아보는 방법으로는 훈련군이 한 개의 연구로 구성된 경우와 두 개 연구로 구성된 경우의 정분류율을 비교하였다.

정분류율의 정확한 예측을 위하여 훈련군과 시험군을 각 3등분한 후 훈련군에 속한 표본의 2/3만 이용하여 특이발현 유전자를 검색하고 이를 바탕으로 시험군 표본의 1/3의 표현형을 예측하는 방법을 전체적으로 3번 반복하였다.

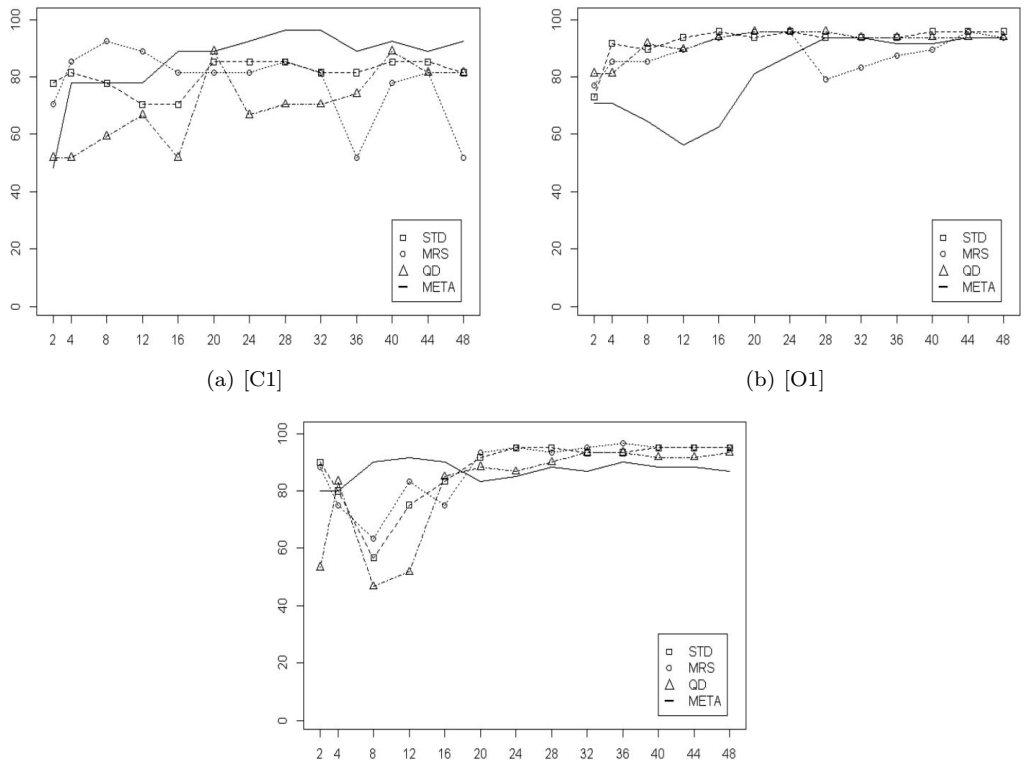
판별방법으로는 많이 사용되는 선형판별분석(linear discriminant analysis), support vector machine, k-nearest neighbor와 신경망(neural network)을 이용한 기법을 사용하였다. 판별방법에 따라 예측 정확성에 약간의 차이가 있었으나 본 논문의 관심은 통합 방법의 비교에 있기 때문에 훈련군과 시험군의 플랫폼 형태에 상관없이 거의 모든 경우에 예측 정분류율이 높게 나타났던 신경망 방법을 이용하여 예측 정확성을 구하였고 이들의 크기를 비교하여 각 통합 방법의 우수성과 통합의 필요성을 판단하였다.

4.1. DLBCL 자료 분석

악성림프종의 하나인 DLBCL은 종양 형성 구조에 따라 치료법에 대한 반응이 다르고 생존 시간에 큰 차이를 보인다. Alizadeh 등 (2000)은 DLBCL 환자들의 유전자 발현을 클러스터 분석하여 germinal center B-cell(GCB)과 activated B-Cell(ABC)의 두 개의 새로운 아형이 존재함을 발견하였고 GCB형 환자가 5년 이상 생존할 가능성이 60% 이상인 반면, ABC형 환자의 45%가 3년 이내 사망함을 보였다. 이러한 차이로 ABC와 GCB간의 유전자적인 차이에 대해 정확히 알아보고 그에 알맞은 치료법을 개발하기 위한 여러 연구가 진행되고 있다. 본 논문에선 Alizadeh 등 (2000), Williams 등 (2010)과 Shaknovich 등 (2010)에서 사용하였던 연구 자료들을 이용하였다 (Table 4.1 참조). 이들 연구에 사용된 플랫폼은 cDNA칩과 affymetrix사의 oligonucleotide 칩으로 구성되어 있으며 GEO에서 다운받을 수 있다.

각 연구에서 결측치가 20% 이상되는 유전자는 제거하고 entrez gene ID를 이용하여 각 유전자의 ID mapping을 한 후 중복된 유전자들은 중위수를 대푯값으로 하였다. 세 개의 연구에 공통으로 존재하는 유전자의 수는 2,590개였으며 이들을 중심으로 자료 분석을 시행하였다.

Table 4.1의 세 개의 연구를 각각 시험군으로, 나머지 두 연구를 통합하거나 단독으로 훈련군으로 택하는 방법을 사용하였다. 훈련군의 두 연구 자료를 순위기반 중위수 변환, 분위수를 이용한 이산화와 표준화를 각각 이용하여 통합하거나 두 자료의 분석 결과에 메타분석을 실시하여 특이발현 유전자들을 찾아



(a) [C1] (b) [O1]

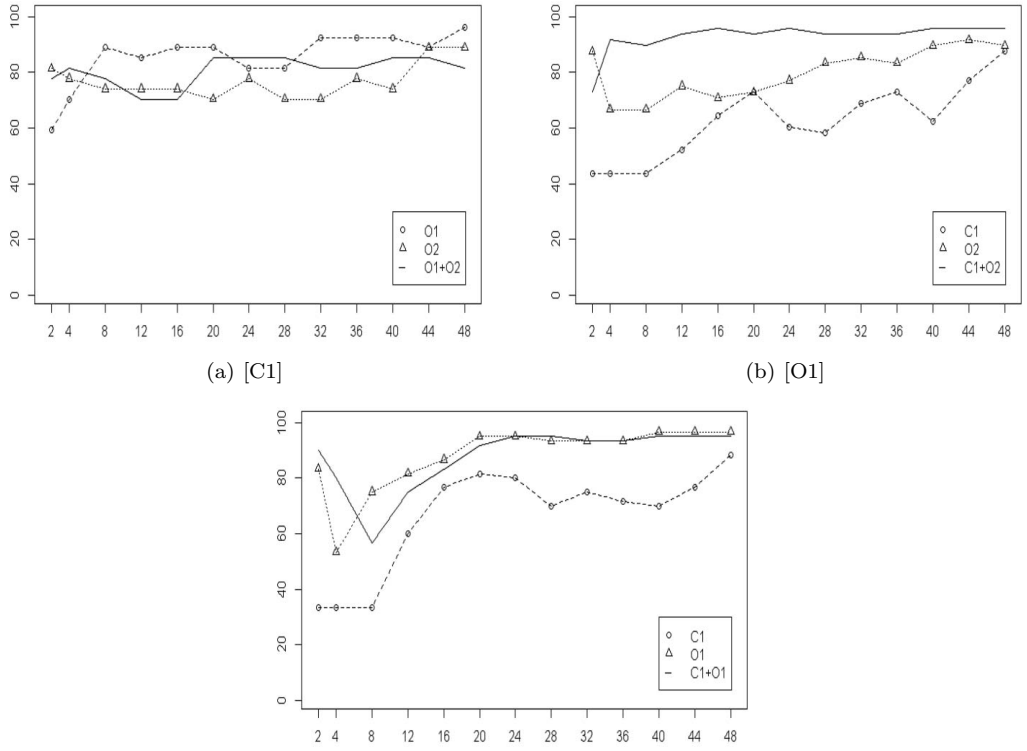
(c) [O2]

Figure 4.1. Plots of accuracies for classification of DLBCL depending on 4 different integrating methods. Each plot uses a different test set([C1], [O1], [O2]). Vertical axis corresponds to the classification accuracy and horizontal axis corresponds to the number of differentially expressed genes used.

낸 후 이를 대상으로 신경망 판별방법을 사용하여 시험군의 표본들의 표현형을 예측하였다.

Figure 4.1은 각 지정된 시험군에 대하여 두 개의 연구를 네 가지 방법으로 각각 통합한 훈련군에서 구한 특이발현 유전자 수를 늘려 가면서 시험군에 속한 표본의 표현형을 예측해 봄으로써 통합 방법에 따른 정분류율을 비교한 것이다. 전체적으로 중위수 변환이나 이산화 방법이 표준화 방법보다 정분류율의 안정성이 떨어졌으며 메타분석은 Figure 4.1(b)에서 특이발현 유전자 수가 적으면 정분류율이 낮아지는 현상을 보였다. Figure 4.1(a)의 훈련군인 [O1]과 [O2]는 실제로 같은 칩을 사용하여 유전자 발현값의 분포 차이가 거의 없는 경우로서 네 가지 방법의 정분류율 차이는 자료의 단순화가 표현형의 예측에 미치는 영향을 반영한 것으로 생각한다

훈련군의 표본 수를 늘리기 위해 플랫폼이나 실험 배경이 다른 연구들을 합치는 것이 새로운 표본의 표현형을 예측하는 데 도움이 되는지 알아보기 위하여 세 연구를 각각 돌아가며 시험군으로 하고 나머지 두 연구는 각각, 그리고 표준화 방법으로 통합하여 훈련군으로 사용하였을 때 예측 정분류율의 차이를 비교하였다 (Figure 4.2). DLBCL의 세 연구 중 [C1]이 시험군으로 사용될 때 (Figure 4.2(a))는 특별할 문제가 없어 보였지만 단독 훈련군일 때 (Figure 4.2(b), (c))는 정분류율이 많이 낮은 것을 볼 수 있다. 이러한 낮은 정분류율은 훈련군의 표본 수가 작거나 정보의 질이 좋지 않을 때 생길 수 있는 현상인



(a) [C1] (b) [O1]

(c) [O2]

Figure 4.2. Comparison of DLBCL classification accuracies before and after integrating different data sets. Each plot uses a different test set([C1], [O1], [O2]). Vertical axis represents to classification accuracy and horizontal axis represents to the number of differentially expressed genes used.

데 Figure 4.2(b)에서 [O2]와의 통합으로 정분류율이 많이 높아진 것을 보면 표본의 수가 27 밖에 되지 않았기 때문인 것 같다. Figure 4.2(a)에서는 훈련군으로서 [O1]과 [O2]의 역할을 비교해 볼 수 있는데 [O1]에서의 특이발현 유전자 검색이 [O2]보다 더 정확한 판단에 도움이 됨을 보이고 있고, 같은 맥락으로 [O1]을 이용하여 [O2] (Figure 4.2(c))를 예측할 때의 정분류율이 [O2]를 이용하여 [O1]을 예측할 때 (Figure 4.2(b))보다 높다는 것을 볼 수 있다.

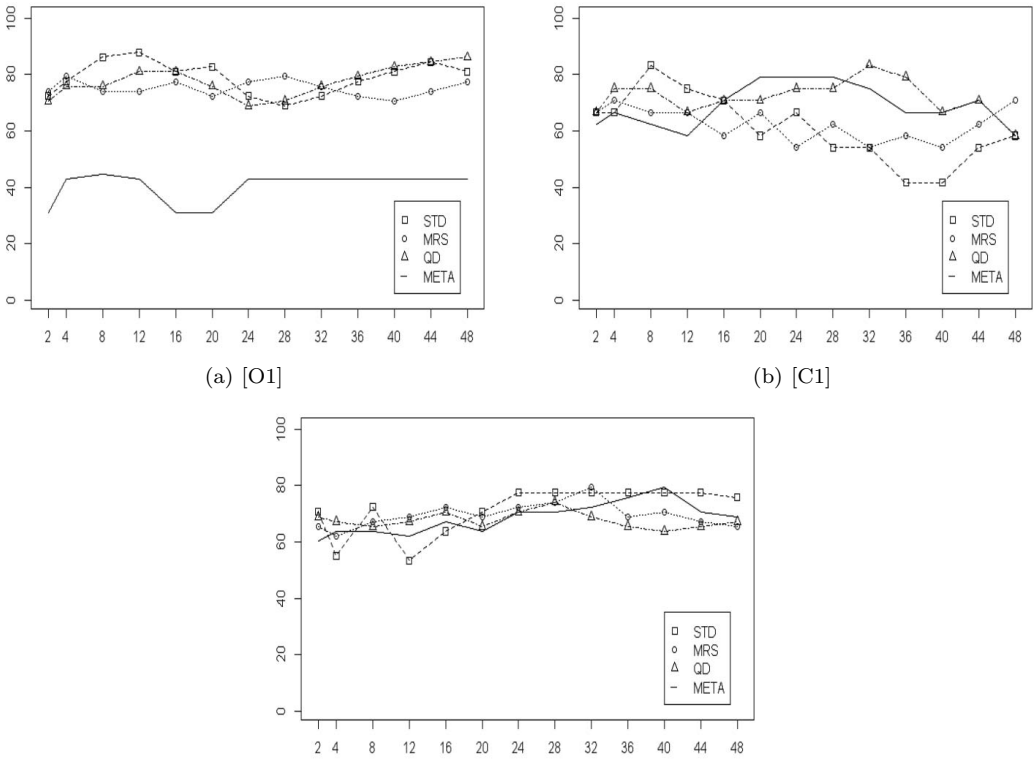
서로 플랫폼이 다른 연구라도 통합을 함으로서 정분류율을 높이는데 기여할 수 있지만 (Figure 4.2(b)) 이미 정분류율이 충분히 높은 경우에는 표본수가 많아져도 더 이상 정분류율을 높이는데 한계가 있음을 볼 수 있다 (Figure 4.2(c)).

4.2. 폐암 자료

전체 폐암 중 80% 이상 차지하는 비소세포폐암(non-small cell lung cancer)은 편평세포암(squamous cell carcinoma; SCC), 선암(adenocarcinoma; AD) 그리고 대세포암(large cell carcinoma) 등으로 구분된다. 본 연구에서는 GEO를 통하여 SCC와 AD를 분류하는 최근 연구를 검색한 결과, 올리고칩을 사용한 한 개의 연구와 서로 다른 종류의 cDNA칩을 사용한 두 개의 연구 결과를 찾았다 (Table 4.2).

Table 4.2. Details of lung cancer microarray studies

Name	Study	Platform	Samples	Probes	GEO
[C1]	Kuner *et al.* (2009)	Oligo	40 AD, 18 SCC	54675	GSE10245
[O1]	Hu *et al.* (data public on 2011)	cDNA	16 AD, 8 SCC	9984	GSE29827
[O2]	Newnham *et al.* (2011)	cDNA	33 AD, 25 SCC	11066	GSE25326



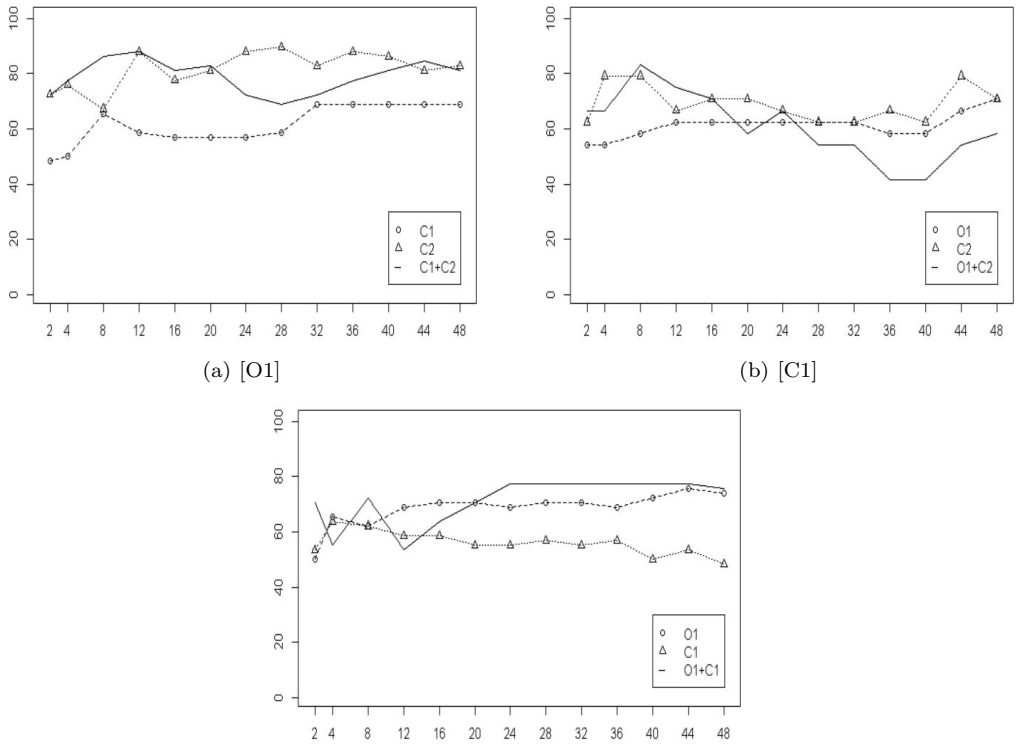
(a) [O1] (b) [C1]

(c) [C2]

Figure 4.3. Plots of accuracies for classification of lung cancer depending on the 4 different integrating methods. Each plot uses a different test set([O1], [C1], [C2]). Vertical axis corresponds to the classification accuracy and horizontal axis corresponds to the number of genes.

각 연구에서 결측치가 20% 이상 되는 유전자는 제거하고 entrez gene ID를 이용하여 각 유전자의 ID mapping을 한 후 중복된 유전자들은 중위수를 대푯값으로 하였다. 세 개의 연구에 공통으로 존재하는 3,672개 유전자를 중심으로 통합 자료 분석을 하였다.

Figure 4.3는 DLBCL 자료 분석과 마찬가지로 세 개의 연구를 각각 시험군으로, 나머지 두 연구를 네 가지 방법에 따라 하나의 훈련군으로 통합하고 여기서 얻은 특이발현 유전자 수를 늘려 가면서 통합 방법에 따른 정분류율을 비교한 것이다. 이 그림에서의 특이점은 네 가지 방법 중 어느 방법이 뚜렷이 우수하다고 할 수는 없지만, Figure 4.3(a)에서 메타분석이 다른 방법들에 비해 정분류율이 많이 낮고 Figure 4.3(b)의 정분류율들은 Figure 4.3(a)와 (c)에 비해 낮고 덜 안정적이라는 것이다. 이에 대한 원인은 연구 [C1]의 문제라 생각된다. SAM을 이용하여 개별연구 분석을 하였을 때 q-value < 0.01을 기



(a) [O1] (b) [C1]

(c) [C2]

Figure 4.4. Comparison of lung cancer classification accuracies before and after integrating different data sets. Each plot uses a different test set([O1], [C1], [C2]). Vertical axis represents to classification accuracy and horizontal axis represents to the number of genes.

준으로 유의한 특이발현 유전자 수가 [O1]은 833개, [C2]는 8개인데 비해 [C1]에는 존재하지 않았다. [C1]의 표본들이 AD와 SCC 사이의 차이를 명확히 보이지 않기 때문에 [C1]을 시험군으로 삼은 Figure 4.3(b)의 정분류율이 낮아진 것 같다. 네 가지 방법 중 메타분석은 통합 대상 연구들이 같은 비중으로 통합되는 데 비해 자료 변환을 이용하는 나머지 세 방법은 각 연구의 표본 수가 통합의 가중치로 작용한다. 이러한 특징은 표본 수가 작고 표현형에 따른 차이가 명확하지 않은 [C1]이 다른 연구와 통합할 때 통합 방법과 대상에 따라 정분류율이 달라지리라 예상할 수 있었는데 특히 Figure 4.3(a)의 메타분석을 이용한 [C2]와의 통합 결과가 좋지 않았다. Figure 4.3(c)에서는 [O1]에 지배적인 특이발현 유전자들이 존재하였기 때문에 [C1]이 큰 영향을 미치지 않은 것 같다.

Figure 4.4에서는 하나의 시험군에 대하여 나머지 두 연구를 각각 또는 표준화 방법을 이용하여 통합한 후의 정분류율을 비교해 보았다. 서로 다른 시험군을 사용한 세 개의 그래프에서 모두 다른 유형의 통합 효과를 보이고 있고, 공통적으로는 [C1]이 훈련군인 경우는 다른 단일 연구 자료나 통합된 연구 자료를 훈련군으로 하였을 때 비해서 정분류율이 낮음을 볼 수 있다. Figure 4.4(a)에서는 [C1] 연구 결과가 명확하지 못하여 통합 결과의 정분류율이 [C1]만을 이용한 단일연구의 정분류율보다는 높아졌지만 [C2]만을 이용하였을 때보다 더 높아지지는 않았다. Figure 4.4(b)에서는 통합을 함으로써 정분류율이 더 떨어지는 기현상을 보였지만 이는 [C1]이 시험군이기 때문이라 여겨진다.

5. 결론

마이크로어레이 실험의 표본 수가 대체로 작은 단점을 보완하기 위하여 실험 목적이 같은 연구들의 표본들을 통합하여 분석하는 것을 다룬 연구가 많이 발표되었다. 그런데 대부분의 연구가 유전자 발현값의 형태가 다른 표본들을 통합하는 방법을 개발하여 나름대로 그 효율성을 제시하는데 그쳤고 여러 통합 방법들을 총체적으로 비교한 연구 보고는 매우 미약하였다. 본 연구에서는 여러 통합 방법의 상대적 비교를 목적으로 공용 데이터 저장소에서 자료를 다운받아 통합 분석을 시도하였는데 많은 문제점이 있었다.

우선 저장소에서 동일한 실험 목적으로 명시된 마이크로어레이 자료는 많이 찾을 수 있었지만 그들의 실험 환경, 표본의 조건 등이 조금씩 차이가 있거나 파악하기 어려워 통합 대상 여부를 가리기가 어려웠다. 실제로 다른 논문에서 이용한 예제 자료들도 몇 가지 되지 않았다. 또한, 통합 연구에서는 통합 대상의 모든 연구에 공통으로 존재하는 유전자만 사용하기 때문에 쓸모없이 버려지는 정보량이 많다는 것이다. 특히 올리고칩과 cDNA칩의 자료를 합할 때 결측자료를 제거하고 나면 올리고칩의 5% 정도만 분석 대상이 되는 경우도 있었다.

DLBCL과 폐암의 실제 자료들을 사용하여 서로 다른 연구를 통합하는 방법과 통합의 필요성을 검토한 결과, 이산화 방법은 유전자 발현값을 7분위수로 단순화시켜 분산이 0이 되는 경우도 간혹 발생하여 t 검정으로 특이발현 유전자를 검색할 수 없었으며, 동일한 플랫폼 자료들의 병합에 있어서는 다른 방법들에 비해 정분류율의 안정성이 떨어짐을 관찰할 수 있었다. 중위수 변환방법은 시험군에 속한 표본을 변환하기 위하여 훈련군의 모든 유전자 정보가 다 필요한 불편함이 있었다. 메타분석은 각 연구의 표본 수가 통합의 가중치로 작용하는 다른 방법들과는 달리 표본 수 상관없이 각 연구 결과가 동일한 비중으로 통합되기 때문에 표본 수가 작거나 실험의 질이 낮은 연구를 통합할 때 주의하여야 하고, 특히 표본수의 차이가 큰 연구들의 통합에는 적당하지 않음을 볼 수 있었다. 이러한 문제점들을 제외하고는 방법에 따른 효율성의 차이가 뚜렷하지 않았고 사용자 측면에서 보면 표준화 방법이 각 표본의 정보량만 이용하기 때문에 간편하다는 장점이 있다.

실험 목적이 같은 연구들을 통합하는 것이 정분류율을 높이는데 기여하는 것인지에 대하여는 무조건 표본의 수를 늘리는 것이 도움 되는 것도 아니고 플랫폼이 같은 연구에 비해 다른 플랫폼의 연구가 무조건 불리한 것도 아니라고 판단된다. 여러 연구를 병합할 때 동일한 플랫폼을 사용한 연구들의 병합이 다른 플랫폼 연구들의 병합보다 결과가 더 좋다. 그러나 표본의 수가 크지 않아도 이미 분류 정확성이 높은 경우는 표본수가 더 커져도 정확성이 더 높아질 여지가 많지 않기 때문에 병합의 효과가 크지 않다. 그러나 표본의 수가 적은 경우는 다른 플랫폼의 연구들의 병합도 크게 도움이 된다. 본 논문에 발표한 자료 외에도 몇 가지 pilot 연구를 한 결과, 통합 대상 연구 간에 공통 유전자가 많고 각 연구의 실험들이 올바른 관리 아래에서 이루어져서 단일 연구 분석에서도 표현형에 따라 차이가 뚜렷한 특이발현 유전자들이 존재할 때 통합의 효과가 나타남을 볼 수 있었다.

기존에 사용되는 분류자(classifier)들의 대부분은 훈련군과 시험군의 관찰값 분포가 동일할 때 사용할 수 있는 것인데 서로 다른 플랫폼의 자료를 쉽게 통합 분석하기 위해서는 유전자의 관찰값 대신 순위나 평균과 비교한 대소 관계 등을 이용한 분류자를 개발하는 것이 필요하겠다.

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J. Jr, Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. and Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, **403**, 503–511.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data, *Nature Genetics*, **29**, 365–371.

Campain, A. and Yang, Y. H. (2010). Comparison study of microarray meta-analysis methods, *BMC Bioinformatics*, **11**, 408.

Diehn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J. C., Hernandez-Boussard, T., Rees, C. A., Cherry, J. M., Botstein, D., Brown, P. O. and Alizadeh, A. A. (2003). SOURCE: A unified genomic resource of functional annotations, ontologies, and gene expression data, *Nucleic Acids Research*, **31**, 219–223.

Fisher, R. A. (1932). *Statistical Methods for Research Workers*, 4ed. Oliver and Boyd, Edinburgh.

Good, I. J. (1955). On the weighted combination of significance tests, *Journal of Royal Statistical Society*, **2**, 264–265.

Hong, F. and Breitling, R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments, *Bioinformatics*, **24**, 374–382.

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources, *Nature Protocols*, **4**, 44–57.

Kim, K. Y., Ki, D., Jeung, H. C., Chung, H. C. and Rha, S. Y. (2008). Improving the prediction accuracy in classification using the combined data sets by ranks of gene expressions, *BMC Bioinformatics*, **9**, 283.

Kuner, R., Muley, T., Meister, M., Ruschhaupt, M., Buness, A., Xu, E. C., Schnabel, P., Warth, A., Poustka, A., Sultmann, H. and Hoffmann, H. (2009). Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes, *Lung Cancer*, **63**, 32–38.

Kuo, W. P., Jenssen, T. K., Butte, A. J., Ohno-Machado, L. and Kohane, I. S. (2002). Analysis of matched mRNA measurements from two different microarray technologies, *Bioinformatics*, **18**, 405–412.

Larkin, J. E., Frank, B. C., Gavras, H., Sultana, R. and Quackenbush, J. (2005). Independence and reproducibility across microarray platforms, *Nature Methods*, **2**, 337–344.

Liu, H., Hussain, F., Tan, C. L. and Dash, M. (2002). Discretization: An enabling technique, *Data Mining and Knowledge Discovery*, **6**, 393–423.

Newnham, G. M., Conron, M., McLachlan, S., Dobrovic, A., Do, H., Li, J., Opeskin, K., Thompson, N., Wright, G. M. and Thomas, D. M. (2011). Integrated mutation, copy number and expression profiling in resectable non-small cell lung cancer, *BMC Cancer*, **7**, 11–93.

Rudy, J. and Valafar, F. (2011). Empirical comparison of cross-platform normalization methods for gene expression data, *BMC Bioinformatics*, **12**, 467.

Shabalin, A., Tjelmeland, H., Fan, C., Perou, C. and Nobel, A. (2008). Merging two gene-expression studies via cross-platform normalization, *Bioinformatics*, **24**, 1154–1160.

Shaknovich, R., Geng, H., Johnson, N. A., Tsikitas, L., Cerchietti, L., Greally, J. M., Gascoyne, R. D., Elemento, O. and Melnick, A. (2010). DNA methylation signatures define molecular subtypes of diffuse large B-cell lymphoma, *Blood*, **116**, e81–e89.

Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., Lee, K. Y., Luo, Y., Sun, Y. A., Willey, J. C., Setterquist, R. A., Fischer, G. M., Tong, W., Dragan, Y. P., Dix, D. J., Frueh, F. W., Goodsaid, F. M., Herman, D., Jensen, R. V., Johnson, C. D., Lobenhofer, E. K., Puri, R. K., Schrf, U., Thierry-Mieg, J., Wang, C., Wilson, M., Wolber, P. K., Zhang, L., Amur, S., Bao, W., Barbacioru, C. C., Lucas, A. B., Bertholet, V., Boysen, C., Bromley, B., Brown, D., Brunner, A., Canales, R., Cao, X. M., Cebula, T. A., Chen, J. J., Cheng, J., Chu, T. M., Chudin, E., Corson, J., Corton, J. C., Croner, L. J., Davies, C., Davison, T. S., Delenstarr, G., Deng, X., Dorris, D., Eklund, A. C., Fan, X. H., Fang, H., Fulmer-Smentek, S., Fuscoe, J. C., Gallagher, K., Ge, W., Guo, L., Guo, X., Hager, J., Haje, P. K., Han, J., Han, T., Harbottle, H. C., Harris, S. C., Hatchwell, E., Hauser, C. A., Hester, S., Hong, H., Hurban, P., Jackson, S. A., Ji, H., Knight, C. R., Kuo, W. P., LeClerc, J. E., Levy, S., Li, Q. Z., Liu, C., Liu, Y., Lombardi, M. J., Ma, Y., Magnuson, S. R., Maqsodi, B., McDaniel, T., Mei, N., Myklebost, O., Ning, B., Novoradovskaya, N., Orr, M. S., Osborn, T. W., Papallo, A., Patterson, T. A., Perkins, R. G., Peters, E. H., Peterson, R., Philips, K. L., Pine, P. S., Pusztai, L., Qian, F., Ren, H., Rosen, M., Rosenzweig, B. A., Samaha, R. R., Schena, M., Schroth, G. P., Shchegrova, S., Smith, D. D., Staedtler, F., Su, Z., Sun, H., Szallasi, Z., Tezak, Z., Thierry-Mieg, D., Thompson, K. L., Tikhonova, I., Turpaz, Y., Vallanat, B., Van, C., Walker, S. J., Wang, S. J., Wang, Y., Wolfinger, R., Wong, A., Wu,

J., Xiao, C., Xie, Q., Xu, J., Yang, W., Zhang, L., Zhong, S., Zong, Y. and Slikker, W. Jr. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements, *Nature Biotechnology*, **24**, 1151–1161.

Stec, J., Wang, J., Coombes, K., Ayers, M., Hoersch, S., Gold, D. L., Ross, J. S., Hess, K. R., Tirrell, S., Linette, G., Hortobagyi, G. N., Fraser Symmans, W. and Pusztai, L. (2005). Comparison of the predictive accuracy of DNA array based multigene classifiers across cDNA arrays and Affymetrix Gene Chips, *Journal of Molecualr Diagnosis*, **7**, 357–367.

Tan, P. K., Downey, T. J., Spitznagel, E. L. Jr, Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M. and Cam, M. C. (2003). Evaluation of gene expression measurements from commercial microarray platforms, *Nucleic Acids Research*, **31**, 5676–5684.

Walker, W. L., Liao, I. H., Gilbert, D. L., Wong, B., Pollard, K. S., McCulloch, C. E., Lit, L. and Sharp, F. R. (2008). Empirical Bayes accommodation of batch-effects in microarray data using identical replicate reference samples: application to RNA expression profiling of blood from Duchenne muscular dystrophy patients, *BMC Genomics*, **9**, 494.

Warnat, P., Eils, R. and Brors, B. (2005). Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes, *BMC Bioinformatics*, **6**, 265.

Williams, P. M., Li, R., Johnson, N. A., Wright, G., Heath, J. D. and Gascoyne, R. D. (2010). A novel method of amplification of FFPET-derived RNA enables accurate disease classification with microarrays, *Journal of Molecular Diagnosis*, **5**, 680–686.

Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C. L., Haase, J., Janes, J., Huss, J. W. III and Su, A. I. (2009). BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources, *Genome Biology*, **10**, R130.

서로 다른 플랫폼의 마이크로어레이 연구 통합 분석

이장미[a] · 이선호[a,1]

[a]세종대학교 수학통계학부

(2013년 1월 7일 접수, 2013년 3월 26일 수정, 2013년 3월 26일 채택)

요 약

마이크로어레이 실험의 특성상 표본의 수가 많지 않는 단점을 보완하고 분석 결과를 일반화하기 위하여 공개 저장소에 축적된 자료 중에 연구 목적이 동일한 여러 연구들을 통합하여 분석하려는 시도가 활발하다. 그러나 실험에서 사용한 플랫폼이 서로 다른 경우에는 유전자 관찰값의 분포가 달라지기 때문에 통합이 어렵고 최상의 통합 방법이 제시되어 있지 않다. 본 논문에서는 순위 기반 중위수, 분위수 이산화와 표준화를 각각 이용하여 변환한 자료값을 직접 합치거나 메타분석을 하여 연구 결과를 합치는 방법을 알아 보았다. 또한 GEO에서 다운받은 실제 자료들을 이용하여 네 가지 방법의 장단점과 효과를 비교하였고 서로 다른 연구 자료를 통합하는 것의 영향을 알아보았다.

주요용어: 마이크로어레이 실험, 서로 다른 플랫폼, 통합 분석, 순위 기반 중위수 변환, 분위수 기반 이산화, 표준화, 메타분석.

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단 기초연구사업의 지원을 받아 수행된 연구임(No. 2012-0004330).

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