논문 2010-47CI-5-3

시간열 마이크로어레이 데이터를 이용한 질병 관련 유의한 패스웨이 유전자 집합의 검출

(A Method of Identifying Disease-related Significant Pathways Using Time-Series Microarray Data)

김 재 영*, 신 미 영**

(Jaeyoung Kim and Miyoung Shin)

요 약

최근 특정 질병의 진단이나 예후 예측을 위해 마이크로어레이 실험 데이터를 이용한 질병 관련 바이오마커 검출 연구가 활발히 진행되고 있다. 특히 정상인에 비해 질병 환자군에서 특이하게 발현되는 개별 유전자를 바이오 마커로 이용하는 기존의 방식과는 달리 동일한 생물학적 패스웨이에 관여하는 유전자 집합의 변화를 분석하여 특이하게 발현되는 패스웨이 유전자 집합을 바이오 마커로 사용하는 유전자 집합 분석(Gene-set analysis) 연구가 주목받고 있다. 본 논문에서는 다양한 실험 조건 요인을 가지는 시간열 마이크로어레이 실험 데이터를 이용한 유의한 패스웨이 유전자 집합을 검출하는 방법에 대해 제안한다. 시간열 마이크로어레이 데이터을 이용하여 유전자 집합 분석을 수행하기 위해서는 시간에 따른 유전자 발현값의 변화에 따라 개별 유전자의 유의성을 나타내는 스코어를 maSigPro (microarray Significant Profiles)를 이용하여 계산한 후, 이를 기반으로 전체 유전자의 순위를 결정하여 후보 유전자 집합에 대한 유의성 검증을 윌콕슨 순위합 검증을 통해 수행한다. 후보 유전자 집합의 생성을 위해서는 MSigDB (Molecular Signatures Database)의 패스웨이 정보를 이용하였으며, 본 논문에서 제안한 방법의 검증을 위해 공개된 전립선 암 관련 시간열 마이크로어레이 실험 데이터에 적용한 결과 실제로 전립선암과 관련된 것으로 밝혀진 7개의 패스웨이 중 6개의 패스웨이를 정확하게 검출할 수 있었다.

Abstract

Recently the study of identifying bio-markers for disease diagnosis and prognosis has been actively performed. In particular, lots of attentions have been paid to the finding of pathway gene-sets differentially expressed in disease patients rather than the finding of individual gene markers. In this paper we propose a novel method to identify disease-related pathway gene-sets based on time-series microarray data. For this purpose, we firstly compute individual gene scores by the using maSigPro (microarray Significant Profiles) and then arrange all the genes in the decreasing order of the corresponding gene scores. The rank of each gene in the entire list is used to evaluate the statistical significance of candidate gene-sets with Wilcoxson rank sum test. For the generation of candidate gene-sets, MSigDB (Molecular Signatures Database) pathway information has been employed. The experiment was conducted with prostate cancer time-series microarray data and the results showed the usefulness of the proposed method by correctly identifying 6 out of 7 biological pathways already known as being actually related to prostate cancer.

Keywords: Gene-set analysis, Significant gene-set, Time-course microarray data-set, Gene ranking, Pathway

* 학생회원, 경북대학교 전자전기컴퓨터학부
(Graduate School of Electrical Engineering and Computer Science, Kyungpook National University)
** 평생회원-교신저자, 경북대학교 IT대학 전자공학부
(College of IT Engineering, Kyungpook National University)
※ 이 논문은 2008년도 정부(교육과학기술부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임
(KRF-2008-331-D00558)
접수일자: 2010년4월28일, 수정완료일: 2010년8월31일

Ⅰ. 서 론

최근 특정 질병의 진단이나 예후 예측을 위해 마이크로어레이 실험 데이터를 이용한 질병 관련 바이오마커 검출 연구가 활발히 진행되고 있다. 특히 정상인에 비해 질병 환자군에서 특이하게 발현되는 개별 유전자를 바이오 마커로 이용하는 기존의 방식과는 달리, 동일한 생물학적 패스웨이에 관여하는 유전자 집합의 변화를 분석[1]하여 특이하게 발현되는 패스웨이 유전자 집합을 바이오 마커[2]로 사용하는 유전자 집합 분석(Gene-set analysis)[1, 3] 연구가 주목받고 있다. 지금까지 발표된 유전자 집합 분석 연구들은 주로 두 개의 샘플 클래스(실험군과 대조군)로 이루어진 DNA 마이크로어레이 실험 데이터를 기반으로 하고 있다. 그리하여, 두 샘플 클래스 간에 유의한 발현 차이를 나타내는 유전자 집합을 검출하기 위한 방법들이 주를 이루고 있으며, 이러한 방법으로 GSEA[1], PAGE[4] 등이 제안된 바 있다.

한편, 최근에는 암과 같은 특정 질병과 관련된 다양한 형태의 시간열 마이크로어레이 실험 데이터들이 구축되어 Gene Express Omnibus[5] 등과 같은 공개 데이터베이스를 통해 소개되고 있다. 이러한 데이터들은 특정 조건하에서 시간 경과에 따른 환자군과 정상인 샘플에서의 유전자 발현값의 변화에 관한 반복 실험 결과를 포함하고 있다[6]. 현재까지 시간열 마이크로어레이 실험 데이터를 기반으로 개별 유전자의 유의성을 판단하는 방법들인 EDGE[7], BATS[8], TimeCourse[9] 등이 제안된 바는 있다. 하지만 이러한 시간열 마이크로어레이 실험 데이터들을 기반으로 질병과 관련하여 특이하게 발현되는 유전자 집합을 검출하는 연구는 아직까지 미비한 상황이다.

본 논문에서는 시간열 마이크로어레이 실험 데이터를 기반으로 특정 질병과 관련하여 유의하게 발현되는 패스웨이 유전자 집합을 검출하기 위한 방법을 제안하고자 한다. 이를 위해 MSigDB (Molecular Signatures Database)[1, 10]의 패스웨이 정보를 사전정보로 이용하였으며, 제안한 방법의 실험적 검증을 위해 2007년도에 발표된 Wang[11]의 전립선암 관련 ChIP-on-chip 실험 데이터를 이용하여 분석 실험을 진행하였다.

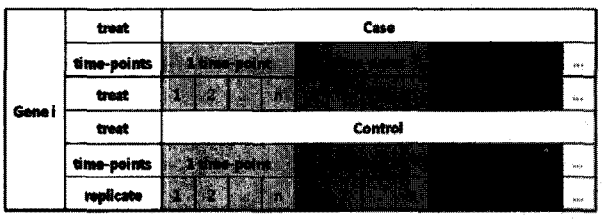
본 논문의 구성은 다음과 같다. 제 Ⅱ장에서는 시간열 마이크로어레이 실험데이터의 특징을 살펴보고, 시간열 데이터에 기반한 각 유전자의 유의성 스코어 계산 방법에 대해 소개한다. 제 Ⅲ장에서는 후보 패스웨이 유전자 집합 생성 방법과 본 논문에서 제안하는 유의한 패스웨이 유전자 집합 검출 방법을 설명한다. 제 Ⅳ장에서는 전립선암 관련 실험 데이터에 대한 유의한 패스웨이 검출 실험결과에 관해 기술하고, 제 Ⅴ장에서는 결론으로 끝을 맺는다.

Ⅱ. 시간열 마이크로어레이 데이터에 기반한 개별 유전자의 유의성 스코어 계산 방법

2.1 시간열 마이크로어레이 데이터의 구성

일반적으로 Gene Express Omnibus 등과 같은 공개 데이터베이스로부터 획득 가능한 시간열 마이크로어레이 실험 데이터들은 아래 그림 1에서와 같은 두 가지 데이터 형태 중의 하나와 같이 구성된다. 첫째는, 그림 1(a)에서와 같이, 하나의 요인(factor)에 대해 시간 경과에 따른 유전자 발현값의 변화를 측정한 것으로 각 시점(time-point)에서 반복실험 결과를 포함하는 종단면 자료(longitudinal data)의 형태를 가진다. 둘째는, 그림 1(b)에서와 같이, 두 개 이상의 요인에 대해 각 요인별로 시간 경과에 따른 유전자 발현값의 변화를 측정한 것으로 각 시점에서 반복 실험 결과를 포함하는 횡단면 자료(cross-sectional data)의 형태를 가진다[6, 9]. 따라서 이러한 시간열 마이크로어레이 데이터를 이용하여 특정 질병에 관련된 유의한 유전자 집합을 검출하기 위해서는 먼저 각 유전자들의 시간열 마이크로어레이 실험 프로파일을 이용하여 개별 유전자의 유의성 판단을

(a)



(b)

그림 1. 시간열 마이크로어레이 실험 데이터의 형태:
(a) Longitudinal 시간열 데이터,
(b) cross-sectional 시간열 데이터

Fig. 1. Time-series microarray data format.
(a) Longitudinal time-series data,
(b) cross-sectional time-series data

위한 유의성 스코어를 계산할 필요가 있다.

2.2 개별 유전자에 대한 유의성 스코어의 계산

시간열 마이크로어레이 실험 데이터에 기반하여 각 유전자들에 대한 유의성 스코어를 계산하기 위해서는 기존의 두 개의 샘플클래스로 이루어진 실험 데이터에 기반한 방법인 SNR[12], t-test[12], SAM[13], fold change[14] 등을 적용할 수 없는 문제가 있다. 이를 위해 최근 시간열 마이크로어레이 실험 데이터 기반의 개별 유전자에 관한 유의성을 판별하는 방법으로 EDGE[7], TimeCourse[9], BATS[8], maSigPro[6] 등이 제안된 바 있다. 이 중에서 maSigPro는 시간열 마이크로어레이 실험 데이터가 서로 다른 시점의 샘플들을 가지거나, 혹은 시점 간격이 동일하지 않는 경우에도 사용할 수 있다는 장점이 있다. 또한, 종단면 및 횡단면 시간열 데이터에 모두 적용할 수 있을 뿐만 아니라 결측치(missing value)가 있거나 실험 샘플의 수가 적은 시간열 데이터에도 적용할 수 있다. 그리하여 본 논문에서는 maSigPro를 이용하여 개별 유전자에 관한 유의성 스코어를 계산하였다.

maSigPro에 의한 각 유전자의 유의성 판단을 위해, 시간열 마이크로어레이 데이터에서의 유전자 발현값들은 시간에 대한 연속적인 값으로 간주한다. 그리하여 각 유전자별로 발현값들에 대한 회귀 모델을 생성함으로써 유의성을 판단한다. 특히 maSigPro는 시간열 유전자 발현 프로파일들의 유의성을 판단하기 위해 다음과 같은 두 단계를 거쳐 유전자의 유의성을 판단한다. 먼저, 횡단면 데이터와 같이 실험군과 대조군 각각에 대해 시점별 발현값이 주어질 경우, 두 클래스에 대한 유의성을 판별하기 위해 각 유전자별로 두 클래스에 대한 회귀모델을 생성하고 회귀모델의 오차값을 이용하여 ANOVA (Analysis of variance) 분석을 통해 유전자의 유의성 스코어를 계산한다. 그리고 두 번째 단계에서는 앞에서 검출된 유전자들의 회귀 모델에 대해 각 시점를 순차적으로 제거하여 생성된 회귀모델과 원래의 유전자 회귀모델들 간의 차이에 대한 통계량을 계산하고 적합도 검증을 통해 각 유전자의 p-value를 계산함으로써 개별 유전자의 유의성을 결정하게 된다.

III. 유의한 패스웨이 유전자 집합의 검출

특정 질병과 관련된 유의한 패스웨이 유전자 집합을 검출하기 위해 본 논문에서는 먼저 실험에 사용된 유전자들이 관여하는 생물학적 패스웨이 정보를 이용하여 후보 패스웨이 유전자 집합을 생성한다. 그리고 이러한 후보 패스웨이 유전자 집합에 속해 있는 개별 유전자들의 유의성 스코어를 이용하여 각 후보 유전자 집합의 유의성을 검증한 후 최종적으로 유의한 패스웨이 유전자 집합을 검출하고자 한다.

3.1 후보 패스웨이 유전자 집합의 생성

시간열 마이크로어레이 실험 데이터에 사용된 유전자들에 관한 후보 유전자 집합들을 생성하기 위하여 공개 생물학 자원인 MSigDB를 사전정보로 사용하여 후보 유전자 집합을 생성하였다. MSigDB는 유전자 온톨로지(Gene-Ontology)나 패스웨이 등과 같은 생물학적인 기능에 따라 유전자들을 분류해 놓은 유전자 집합 정보와 각 유전자 집합에 속한 유전자 ID인 gene-symbol들로 구성되어 있다. 일반적으로 마이크로어레이 실험에 사용된 유전자 칩의 종류에 따라 각 유전자들은 고유의 유전자 번호를 가지고 있다. 이러한 유전자 번호는 유전자를 명시하는 표준 유전자 이름인 gene-symbol로 변환한 후, MSigDB의 유전자 집합 정보를 이용하여 후보 유전자 집합을 생성한다.

본 논문에서는 후보 유전자 집합을 생성하기 위해 특정한 생물학적 기능에 공통으로 관여하는 유전자 그룹을 나타내는 패스웨이 정보를 이용하였다. 즉, 후보 유전자 집합을 생성하기 위해 특정 패스웨이에 속한 유전자 정보들을 이용하여 실험 데이터에 포함된 유전자들의 gene-symbol과 일치하는 유전자들을 추출하고 이들로 구성된 후보 유전자 집합을 생성한다. 이때 후보 유전자 집합에 속해 있는 유전자들의 수가 최소 5개 이상인 유전자 집합[1, 3]만을 유전자-집합 분석에 사용하였다. 후보 유전자 집합에 속해 있는 유전자의 수가 최소 5개 이상인 것만을 사용하는 이유는 유의한 유전자 집합을 검출하기 위한 통계적인 분석을 수행하기 위해서는 샘플의 수가 최소 5개 이상이 되어야 의미 있는 결과를 얻을 수가 있기 때문이다.

3.2 패스웨이 유전자 집합의 유의성 검증

앞서 기술한 바와 같이 MSigDB를 이용하여 생성된 후보 패스웨이 유전자 집합들의 유의성을 판단하기 위해서는 각 유전자 집합에 관한 유의성 검증이 필요하다. 그러나 기존의 두 클래스로 구성된 실험 데이터의

유전자-집합 분석을 위해 사용되었던 PAGE나 GSEA 방법들을 그대로 적용할 수 없는 문제가 있다. 예를 들어, PAGE는 각 유전자들에 대해 fold change를 이용하여 계산된 스코어를 기반으로 유전자를 서열화하고, 후보 유전자 집합에 속한 개별 유전자들의 스코어들로부터 Z-스코어를 계산함으로써 이를 기반으로 유의한 유전자 집합을 검출하였다[4]. 그러나 이 방법은 개별 유전자들에 대한 유의성 스코어가 정규 분포 모델을 가지고 있다고 가정한 후에 분포 모델에 적합한 검증방법을 적용하여 유의한 유전자 집합을 검출하고 있다. 한편, GSEA는 후보 유전자 집합들에 대해 ES (Enrichment Score)값을 계산하고 두 실험 그룹에 대해 순열 검증을 적용하여 유의한 유전자 집합들을 검출한다[1, 3]. 그러나 이 방법은 개별 유전자의 스코어 그래프가 그림 3에서와 같은 형태를 지니고 있을 때만이 적용 가능하다.

본 논문에서 사용한 maSigPro의 경우 각 유전자의 시간열 마이크로어레이 데이터에 대해 계산된 유의성 스코어의 분포는 그림 2와 같다. 그림 2에서 알 수 있듯이 개별 유전자들의 스코어 분포가 정규분포모델을 따른다고 할 수 없다. 실제로 정규성 검증 방법 중의 하나인 Kolmogorov-Smirnov 적합도 검증[15, 16]을 수행한 결과 p-value가 *2.2e-16*로 산출되어 표준 정규 분포를 따르지 않는다고 볼 수 있다. 따라서 PAGE를 이용한 유의 유전자 집합 검증 방법을 적용할 수 없다. 또한, maSigPro를 적용하여 얻은 각 유전자의 유의성 스코어를 정렬한 결과 그림 4와 같으며, 이는 그림 3과는 다른

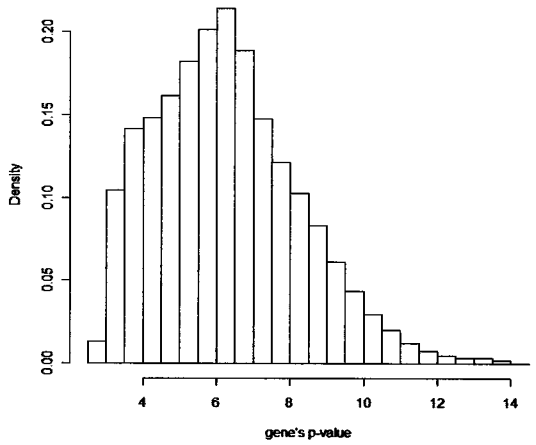


그림 2. Wang의 전립선암 실험 데이터에 maSigPro를 적용하여 획득한 유전자 유의성 스코어들의 분포

Fig. 2. Distribution of gene significance scores for Wang's prostate cancer data.

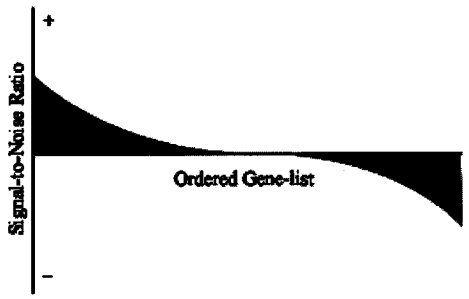


그림 3. GSEA에서 유전자 스코어에 기반한 정렬 그래프의 형태[1]

Fig. 3. A typical form of gene scores graph in GSEA.

형태이다. 따라서 maSigPro를 적용한 유전자 유의성 스코어를 기반으로 유의한 유전자 집합을 검출하기 위해서는 GSEA 방법을 사용할 수 없다.

실제로 다양한 시간열 마이크로어레이 실험 데이터에 여러 가지 유전자 유의성 판단 방법을 적용하여 획득한 유전자 스코어들의 분포 모델을 살펴보면 특정 분포 모델을 가정할 수 없는 경우가 대부분이다. 그리하여 본 논문에서는 유전자 스코어들이 특정한 분포 모델을 따르지 않는다고 가정하고 비모수적 검증 방법[17~18]의 하나인 데이터의 순위를 이용한 윌콕슨 순위합검증 방법[17]을 적용하여 유의한 유전자 집합을 검출하였다. 즉, 앞서 생성된 후보 유전자 집합의 유의성을 검증하기 위해 기존의 방식처럼 유전자 스코어를 이용하지 않고 유전자 스코어에 의한 순위를 이용함으로써 유의한 유전자 집합을 검출하는 것이다. 그림 5는 각 후보 유전자 집합에 속한 유전자들의 순위값(ranking scores)을 어떻게 결정하는지를 나타낸 그림이다. 실험에 사용된 모든 유전자들에 대해 maSigPro를 이용하여 유의성

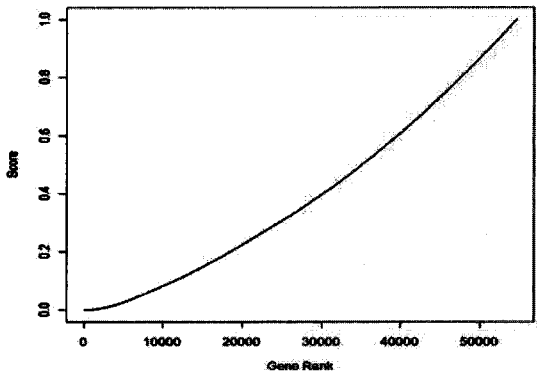


그림 4. Wang의 전립선암 실험 데이터에 maSigPro를 적용하여 획득한 유전자 스코어에 기반한 정렬 그래프

Fig. 4. A graph of gene scores obtained by maSigPro for Wang's prostate cancer data.

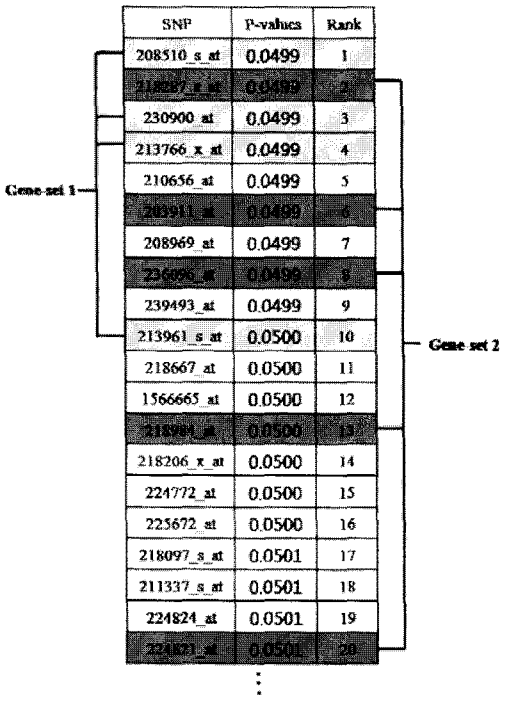


그림 5. 각 후보 유전자 집합에 속한 유전자들의 순위 값 결정

Fig. 5. Determination of ranking scores for individual genes in a gene-set.

스코어를 구하고, 이 스코어에 의해 유전자를 오름차순으로 정렬함으로써 각 유전자의 최종 순위를 결정한다.

이렇게 각 후보 유전자 집합에 속한 유전자들의 순위가 결정되면, 후보 유전자 집합들에 대해 윌콕슨 순위합검증 방법을 이용한 유의성 검증이 수행된다. 즉,각 후보 유전자 집합 S에 속한 유전자들의 순위 R_j들을 이용하여 각 후보 유전자 집합에 대한 순위합 통계량 W_S을 다음과 같이 계산한다.

$$W_S = \sum_{j=1}^{n_S} R_j$$

이 때 n_S는 후보 유전자 집합에 포함된 유전자의 수를 나타낸다. 만약, 실험에 사용된 전체 유전자들의 개수에서 특정 후보 유전자 집합에 포함된 유전자들(n_S)을 제외한 개수를 m_S라 하면, 전체 유전자들의 개수는 m_S+n_S가 된다.

이제 각 후보 유전자 집합의 유의성 검증은 전체 유전자들로부터 생성 가능한 모든 후보 유전자 집합에 대한 순위합 분포를 고려할 때 특정 후보 유전자 집합의 순위합에 대한 유의성을 판단함으로써 결정한다. 실제로 이를 위해서는 실험에 사용된 전체 유전자들의 수가 충분히 많기 때문에 전체 유전자들로부터 생성 가능한 유전자 집합의 순위합 W_{total}에 대한 분포는 순위합들의 평균을 중심으로 표준정규 분포를 따른다고 볼 수 있고, 평균과 분산은 다음과 같이 계산할 수 있다[17, 18].

$$E(W_{total}) = \frac{n_S(m_S + n_S + 1)}{2}$$

$$\mathrm{var}(W_{total}) = \frac{m_S n_S(m_S + n_S + 1)}{12}$$

또한, 전체 유전자들의 순위합 W_{total}의 평균 $E(W_{total})$와 분산 $var(W_{total})$를 이용하여 특정 후보 유전자 집합 S에 대한 Z-스코어 Z_S을 다음과 같이 구할 수 있다.

$$Z_S = \frac{W_S - E(W_{total})}{\sqrt{\mathrm{var}(W_{total})}}$$

그리하여 특정 후보 유전자 집합 S에 대한 유의성은 Z-스코어에 대한 p-value를 구함으로써 최종 결정된다. 본 논문에서는 후보 유전자 집합의 p-value가 0.05이하인 것을 유의한 유전자 집합으로 검출하였다.

Ⅳ. 실 험

본 연구에서 제안한 방법의 유용성을 검증하기 위해 2007년 Wang[11]이 실험한 Androgen(남성호르몬) Receptor 관련 전립선암의 성장[19, 20]에 관한 시간열 마이크로어레이 실험 데이터를 분석하였다. 이 데이터는 54,675 개의 유전자들로 구성되어 있으며, Affymetrix사의 HG-U133 Plus 2 Transcription Factor와 관련된 ChIP-on-chip을 이용한 실험으로 3개의 시점에 대해 3번의 반복 실험을 수행한 종단면 형태의 데이터집합이다. 그림 6은 Wang의 시간열 실험 데이터에 대한 실제 구성 형태를 보여주고 있다.

본 실험에 사용된 시간열 데이터에 대해 maSigPro을 적용하여 각 유전자의 스코어를 계산하고 이를 기반으

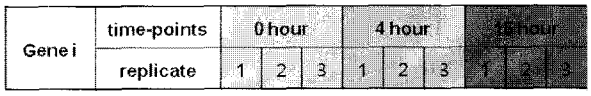


그림 6. Wang의 전립선암관련 시간열 마이크로어레이 실험 데이터의 형태

Fig. 6. Wang's prostate cancer data format.

로 유전자 순위를 결정하였다. 또한, 후보 유전자 집합을 생성하기 위해 MSigDB에서 c2에 해당하는 KEGG pathway 생물학적 자원[21, 22]을 이용하여 54,675 개의 유전자들이 관여하는 전체 패스웨이 중 유전자의 수가 최소 5개 이상을 포함하는 패스웨이들만을 대상으로 유전자 집합 187개를 생성하였다. 이렇게 생성된 각 패스웨이 유전자 집합들에 대해 그림 5에서와 같이 후보 유전자 집합들에 속한 유전자들의 순위를 합산하여 순위합 통계량을 계산하였다. 그리고 월콕슨 순위합검증을 적용하여 유의 수준이 0.05이하(p-value≦0.05)인 유의한 패스웨이 유전자 집합 60개를 최종 검출하였다.

본 논문에서 제안한 방법에 의해 검출된 유의한 패스웨이 유전자 집합 60개가 과연 생물학적으로 의미 있는 결과인지를 검증하기 위해 KEGG 패스웨이 데이터베이스[22]로부터 전립선암과 관련하여 이미 알려진 7개의 패스웨이를 찾아내고 이를 Gold Standard로 활용하였다. 표 1은 KEGG 패스웨이 데이터베이스로부터 찾아낸 전립선암 관련 이미 알려진 패스웨이들을 정리하여 나타낸 것이다.

본 논문에서 제안한 방법을 통해 유의한 유전자 집합으로 검출된 60개 중에서 표 1에 나타난 전립선암 관련 알려진 패스웨이 7개 중 Cytokine-cytokine receptor interaction을 제외한 총 6개의 패스웨이들을 정확히 검출하였다. 표 2는 본 논문에서 제시한 방법을 통해 최종 획득한 전립선암 관련 패스웨이들의 p-value를 나타낸 것이다. 표 2에 따르면, 패스웨이 ID가 hsa04060인 후보 유전자 집합 외에 다른 패스웨이 유전자 집합들의 p-value는 유의 수준 0.05보다 작게 나왔다. 즉, 실제 전립선암과 관련된 패스웨이 7개 중 6개를 유의한 유전자 집합으로 검출할 수 있었다.

반면에 기존의 유전자 집합 분석 방법인 Fisher's exact test의 경우 전립선암 데이터를 maSigPro를 적용하여 2,154개의 유의한 유전자들을 검출하고, Fisher's exact test[12, 15]를 통해 27개의 유의한 유전자 집합을 검출한 후에 결과를 비교하였다. 표 3에서와 같이 7개의 전립선암 관련 패스웨이 중에서 3개만을 정확히 검출하였다.

표 1. KEGG로부터 찾아낸 전립선암 관련 알려진 패스웨이들[22]

Table 1. A list of known pathways related to prostate cancer in KEGG.

Pathway ID	Pathway Name
hsa00150	Androgen and estrogen metabolism
hsa04010	MAPK signaling pathway
hsa04060	Cytokine-cytokine receptor interaction
hsa04110	Cell cycle
hsa04115	p53 signaling pathway
hsa04210	Apoptosis
hsa05215	Prostate cancer

표 2. Wang 데이터에 대한 유전자 집합 분석 결과

Table 2. Result of gene-set analysis for Wang's prostate cancer data.

Pathway gene-set	P-value
Androgen and estrogen metabolism	4.683547e-02
MAPK signaling pathway	4.357878e-03
Cytokine-cytokine receptor interaction	9.973624e-01
Cell cycle	2.556230e-03
p53 signaling pathway	1.768381e-04
Apoptosis	3.061455e-04
Prostate cancer	2.688875e-05

V. 결 론

본 논문에서는 시간열 마이크로어레이 실험데이터와 같이 시간에 따라 유전자 발현값의 변화를 가지는 실험데이터 집합을 기반으로 유의한 유전자 집합을 검출하는 방법을 제안하였다. 기존의 유전자 집합 분석 방법의 경우 두 클래스를 가지는 마이크로어레이 실험데이

표 3. 본 논문에서 제시한 방법과 기존의 Fisher's exact test 결과 비교분석

Table 3. Results of gene-set analysis methods for prostate cancer data sets.

Pathway gene-set	Proposed method	Fisher's exact test[12]
Androgen and estrogen metabolism	●	X
MAPK signaling pathway	●	X
Cytokine-cytokine receptor interaction	X	X
Cell cycle	●	X
p53 signaling pathway	●	●
Apoptosis	●	●
Prostate cancer	●	●

터를 기반으로 유의한 유전자 집합을 검출할 수 있었던 반면에, 시간열 데이터에는 적용하지 못하는 문제가 있다.

본 논문에 제시한 방법을 통해 전립선 암 관련 다양한 실험요인을 가지는 시간열 마이크로어레이 실험 데이터 집합에 적용한 결과 특정 질병이나 약물에 대한 생물학적으로 유의한 유전자 기능 및 유전자들을 검출할 수 있었다. 특히, 기존의 방법에 비해 생물학적으로 좀 더 의미 있는 유전자 집합을 검출할 수 있었다.

참 고 문 헌

[1] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," Proc. Natl Acad Sci USA 102: 15545-50, Sep. 2005.

[2] E. Lee, H.Y. Chuang, J.W. Kim, T. Ideker, D. Lee, "Inferring pathway activity toward precise disease classification," PLoS Comput Biol. 4(11):e1000217, Nov. 2008.

[3] E. Taskesen, "Sub-typing of model organisms based on gene expression data." Bioinformatics technical University of Delft Research Assignment, 2006.

[4] S.Y. Kim, D.J. Volsky, "PAGE: parametric analysis of gene set enrichment," BMC Bioinformatics, 8;6:144, Jun. 2005.

[5] "Gene Expression Omnibus", Available: http://www.ncbi.nlm.nih.gov/geo/

[6] A. Conesa, M.J. Nueda, A. Ferrer, M. Talon, "maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments," Bioinformatics, 1;22(9):1096-102, May 2006.

[7] J.T. Leek, E. Monsen, A.R. Dabney, J.D. Storey, "EDGE: extraction and analysis of differential gene expression," Bioinformatics, 15;22(4):507-8, Feb. 2006.

[8] C. Angelini, L. Cutillo, D. De Canditiis, M. Mutarelli, M. Pensky, "BATS: a Bayesian user-friendly software for analyzing time series microarray experiments,", BMC Bioinformatics, 6;9:415, Oct. 2008.

[9] Y.C. Tai, T.P. Speed, "On Gene Ranking Using Replicated Microarray Time Course Data," Biometrics, ;65(1):40-51, Mar. 2009.

[10] "MSigDB: Molecular Signatures Database", Available: http:// www. broadinstitute.org/ gsea/index.jsp

[11] Q. Wang, W. Li, X.S. Liu, J.S. Carroll, O.A. Jänne, E.K. Keeton, A.M. Chinnaiyan, K.J. Pienta, M. Brown, "A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth," Molecular Cell, 3;27(3):380-92, Aug. 2007.

[12] Zhang, "Advanced analysis of gene expression microarray data," World Scientific, 2006.

[13] S. Baek , H. Moon, H. Ahn, R.L. Kodell, C.J. Lin, J.J. Chen, "Identifying high-dimensional biomarkers for personalized medicine via variable importance ranking," J Biopharm Stat., 18(5):853-68, 2008.

[14] V. Zuber, K. Strimmer, "Gene ranking and biomarker discovery under correlation," Bioinformatics, 15;25(20):2700-7, Oct. 2009.

[15] R. Gentleman, V. Carey, W. Huber, R. Irizarry and S. Dudoit, "Bioinformatics and Computational Biology Solutions Using R and Bioconductor," Springer, 2005.

[16] J. Verzani, "Using R for Introductory Statistics," Chapman & Hall/CRC, Boca Raton, FL, 2005.

[17] 송문섭, 박창순, 이정진, "S-LINK를 이용한 비모수통계학," 자유아카데미, 2003.

[18] R.V. Hogg, A.T. Craig, J. Mckean, "Introduction to Mathematical Statistics, 6th Edition," Pearson Education, 2005.

[19] S. Kudsen, "Cancer Diagnostics with DNA Microarrays," John Wiley & Sons, Inc., 2006.

[20] R.A. Weinberg, "The biology of CANCER," Carland Science, 2007.

[21] Kanehisa M., Goto S., Kawashima S., Nakaya A., "The KEGG databases at GenomeNet," Nucleic Acids Res., 30:42-46, 2002.

[22] "KEGG (Kyoto Encyclopedia of Genes and Genomes) PATHWAY Database", Available: http://www.genome.ad.jp/kegg/pathway.html

[23] Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E.S., Loda, M., Kantoff, P. W., Golub, T. R., Sellers, W. R., "Gene ex-pression correlates of clinical prostate cancer beha-vior." Cancer Cell, vol. 1, no. 2, pp.203-209, Mar. 2002.

저 자 소 개

김 재 영(학생회원)
2006년 위덕대학교 컴퓨터공학과 학사 졸업.
2009년 경북대학교 대학원 정보통신학과 졸업.
2009년 3월~현재 경북대학교 전자전기컴퓨터학부 박사과정.
<주관심분야 : 생물정보학, 데이터 마이닝, 패턴 인식>

신 미 영(평생회원)-교신저자
1991년 연세대학교 전산과학과 학사 졸업.
1993년 연세대학교 전산과학과 석사 졸업.
1998년 미국 Syracuse Univ., EECS Dept. Ph.D.
1999년~2005년 3월 한국전자통신연구원 선임연구원
2005년 4월~현재 경북대학교 IT대학 전자공학부 부교수
<주관심분야 : 패턴인식, 생물정보학>