논문 2008-45CI-4-3

Fisher Criterion을 이용한 Gene Set Enrichment Analysis 기반 유의 유전자 집합의 검출 방법 연구

(Identifying Statistically Significant Gene-Sets by Gene Set Enrichment Analysis Using Fisher Criterion)

김 재 영*, 신 미 영**

(Jae-Young Kim and Miyoung Shin)

요 약

Gene set enrichment analysis (GSEA)는 두 개의 클래스를 가지는 마이크로어레이 실험 데이터 분석을 위해 생물학적 특징을 기반으로 구성된 다양한 유전자-집합 중에서 두 클래스의 발현값들이 통계적으로 중요한 차이를 나타내는 유의한 유전자-집합을 추출하기 위한 분석 방법이다. 특히, 유전자에 대한 다양한 생물학적인 정보를 지닌 유전자 주석 데이터베이스 (Cytogenetic Band, KEGG pathway, Gene Ontology 등)를 이용하여 마이크로어레이 실험에 사용된 전체 유전자 중 특정 기능을 가지는 유전자들을 그룹화하여 다양한 유전자-집합을 발굴하고, 각 유전자-집합 내에서 두 클래스 간에 발현값의 차이를 참조하여 유의한 유전자들을 결정하여, 이를 기반으로 통계적으로 유의한 유전자-집합들을 최종 검출하는 방법이다. 본 논문에서는 GSEA 분석 과정에서 현재 주로 사용되고 있는 signal-to-noise ratio 기반 유전자 서열화(gene ranking) 방법 대신에, Fisher criterion을 이용한 유전자 서열화 방법을 적용함으로써 기존의 GSEA 방법에서 추출하지 못한 생물학적으로 의미 있는 새로운 유의 유전자-집합을 추출하는 방법을 제안하고자 한다. 또한, 제안한 방법의 성능을 고찰하기 위하여 공개된 Leukemia 관련 마이크로어레이 실험 데이터 분석에 적용하였으며, 기존의 알려진 결과와 비교 분석함으로써 제안한 방법의 유용성을 검증하고자 하였다.

Abstract

Gene set enrichment analysis (GSEA) is a computational method to identify statistically significant gene sets showing significant differences between two groups of microarray expression profiles and simultaneously uncover their biological meanings in an elegant way by employing gene annotation databases, such as Cytogenetic Band, KEGG pathways, gene ontology, and etc. For the gene set enrichment analysis, all the genes in a given dataset are first ordered by the signal-to-noise ratio between the groups and then further analyses are proceeded. Despite of its impressive results in several previous studies, however, gene ranking by the signal-to-noise ratio makes it difficult to consider highly up-regulated genes and highly down-regulated genes at the same time as the candidates of significant genes, which possibly reflect certain situations incurred in metabolic and signaling pathways. To deal with this problem, in this article, we investigate the gene set enrichment analysis method with Fisher criterion for gene ranking and also evaluate its effects in Leukemia related pathway analyses.

Keywords: Significant gene-sets, Gene Set Enrichment Analysis, Gene ranking, Fisher Criterion

* 학생회원, 경북대학교 정보통신학과
(Graduate School of Information and Communication Engineering, Kyungpook National University)
** 정회원, 경북대학교 전자전기컴퓨터학부
(School of Electrical Engineering and Computer Science, Kyungpook National University)
※ 본 논문은 2단계 BK21 사업 지원에 의하여 연구되었음.
접수일자: 2008년2월29일, 수정완료일: 2008년7월9일

I. 서 론

DNA 마이크로어레이를 이용한 대량의 유전자에 관한 발현 데이터 분석은 현대 생명 공학 연구에 있어서 매우 중요한 역할을 해왔다. 특히, 최근에는 DNA 마이크로어레이 실험 데이터뿐만 아니라 다양한 사용 가능

한 생물학적 자원들을 함께 분석과정에 이용함으로써 복잡한 생물학적 매커니즘을 규명하기 위한 다양한 연구들이 진행되고 있으며, 많은 의미 있는 결과들을 얻어내고 있다[7, 9, 10]. 전통적으로 DNA 마이크로어레이 데이터 분석에 있어 가장 중요한 문제 중의 하나는 주어진 환경에서 특이하게 발현되는 유의 유전자(differentially expressed genes)를 검출하는 문제이다. 즉, 특정 실험에서 발현량(intensity)이 기준치에 비해 많이 높거나 낮으면서 이러한 현상이 통계적으로 의미가 있는 유전자들을 찾아내는 방법이다. 이러한 유의 유전자 검출 방법의 하나로 최근 많은 주목을 받고 있는 접근 방식은 Gene-Set Enrichment Analysis[1] 방법이다. 이 방법은 두 개의 클래스를 가지는 마이크로어레이 비교 분석 실험에서 특정 생물학적 공통 요소를 지닌 유전자들로 구성된 다양한 유전자-집합(gene-set)을 생성하고, 이 중에서 두 클래스 간에 발현값의 차이가 유의하게 나타나는 유전자-집합을 통계적 분석을 통해 찾아내는 방법이다. 이것은 기존의 유의 유전자 추출 방법들이 다른 생물학적인 정보를 고려하지 않고 수치적인 발현 데이터에만 의존하여 분석함으로써 생물학적 매커니즘을 이해하기 위해 추가적인 별도의 생물학적 분석과정이 필요했던 것과는 달리, 유의 유전자 집합을 추출하기 위한 자체 분석과정에서 생물학적인 정보를 발현 프로파일과 함께 고려하여 유의성을 판단함으로써 이전의 문제를 해결하고자 하였다.

GSEA 분석에서 중요한 단계 중의 하나는 유전자 서열화(gene ranking) 과정으로 현재 이를 위해 주로 사용되고 있는 방법은 signal-to-noise ratio (SNR) 방법[2]이다. 이것은 유전자-집합의 유의성을 판단하기 위해 유전자 발현값이 두 클래스 중 어느 한 클래스에서 상대적으로 매우 높게 나타나는 경우(즉, SNR 값이 양의 영역에 위치한 경우)의 유전자들이거나 혹은 매우 낮게 나타나는 경우(즉, SNR 값이 음의 영역에 위치한 경우)의 유전자들 중에서 어느 한 경우에 속한 유전자들만이 유의한 유전자로 고려되는 문제가 있다. 이러한 문제 때문에 선택되지 않은 다른 영역에 속한 유전자들 중에서 중요한 의미를 지닌 유전자들이 최종적으로 유전자-집합의 유의성을 판단하는 데에 반영되지 못하는 경우가 종종 발생하게 된다.

본 논문에서는 이러한 문제를 해결하기 위하여 Fisher criterion[4~6]에 기반한 유전자 서열화 방법을 적용함으로써 어느 한 클래스에서의 발현값이 다른 클래스에 비해 높거나 낮은 것 중 하나를 선택하기보다는 두 클래스에서 나타나는 발현값의 차이가 통계적으로 의미를 나타내는 유전자들을 대상으로 유의한 유전자-집합을 검출하고자 한다. 데이터 분석 실험을 위해 Golub et al. (1999)의 Leukemia 데이터를 사용하였으며 추출된 유의 유전자 집합의 결과를 기존의 알려진 생물학적 정보와 비교 분석하였다.

본 논문의 구성은 다음과 같다. 제 Ⅱ장에서는 관련 연구로서 GSEA에 관해 요약 설명하고, 제 Ⅲ장에서는 Fisher criterion을 이용한 유전자 서열화 방법에 대해 기술하고, 이의 적용에 따른 GSEA 분석 결과에 관해 논의한다. 제 Ⅳ장에서는 데이터 분석 실험을 통한 검증 결과에 관해 기술하고 제 Ⅴ장에서는 결론 및 토론으로 끝을 맺는다.

Ⅱ. Gene-Set Enrichment Analysis

GSEA는 처리군과 대조군으로 구성된 두 개의 클래스에 속하는 샘플들에 대한 마이크로어레이 발현 프로파일과 관련 유전자의 생물학적인 정보(Cytogenetic Band, KEGG pathway, Gene Ontology 등)를 포함하는 유전자 주석 데이터베이스를 이용하여 유전자의 특성별로 다양한 유전자-집합을 구성하고, 이들 중에서 두 클래스에 속하는 유전자 발현값의 차이가 통계적 유의성을 지니는 유전자-집합 및 유전자를 찾아내는 방법이다. 특히, 스트레스 등의 다양한 자극 여부에 따라 발현값의 차이를 나타내는 유전자들이 특정 기능을 수행하는 유전자-집합에서 유의성을 가지는지를 분석하는데 많이 사용되고 있다[8]. 2005년 발표된 A. Subramanian et al.[1] 연구에서는 GSEA 분석 방법을 이용하여 사람의 혈액 속에 있는 lymphoblastoid 세포에서 남성과 여성 간에 차이를 나타내는 유전자-집합을 추출하거나, NCI-60 암 세포 라인에서의 유전자 발현 데이터를 이용하여 다양한 자극 신호에 따라 유전자 발현을 조절하는 전사인자인 p53의 타깃을 규명하는 데에 적용하였다. 또한, Acute Leukemia와 관련된 연구 및 폐암과 관련한 정상인 유전자와 암에 걸린 유전자들을 비교 분석한 결과를 소개하고 있다[1]. 한편, 2006년 Erdogan Taskesen[2]는 사람의 암 발병과 관련된 중요 유전자 연구를 위해 GSEA를 이용하여 HMEC (Human Mammary Epithelial Cell) cell lines의 패스웨이에 관한 연구를 수행하였다. 또한, 2005년에 Stefano Monti[3]가 발표한 논문에서는 종양 세포에서의 유전자 발현 프로파일을 이용하여 diffuse large B-cell lymphoma

(DLBCL: 범발성 대 B세포 림프종)환자의 생존율과 관련된 유전자들을 GSEA와 패스웨이 정보를 이용하여 유의한 유전자들을 추출한 바 있다.

일반적으로 GSEA 분석을 위해서는 크게 세 단계의 세부 분석 과정을 거치며, 각 단계에서는 유전자-집합의 유의성 판단을 위해 여러 가지 통계학적인 방법들을 사용하고 있다. GSEA 분석의 3단계는 다음과 같이 요약될 수 있다.[1]

[1 단계] 유전자-집합에 대한 Enrichment Score (ES)를 계산하는 단계

- 실험에 사용된 전체 유전자의 마이크로어레이 발현 데이터를 기반으로 두 클래스 간의 발현값 차이에 근거하여 유전자 서열화 기법에서 정한 측정치(metric)에 따라 내림차순으로 전체 유전자 리스트를 정렬한다.
- 유전자 주석 데이터베이스를 이용하여 전체 유전자 리스트 중 특정 기능을 공유하는 유전자들로 구성된 다양한 유전자-집합을 생성한다.
- 정렬된 유전자 리스트와 각각의 유전자-집합들을 이용하여 유전자-집합별 ES 점수를 계산한다.

[2 단계] ES의 통계적 유의성을 추정하는 단계

- 주어진 마이크로어레이 발현 데이터를 클래스 라벨에 대해 k번의 순열(permutation)을 실시하여 k개의 독립적인 유전자 발현 데이터 셀을 구성한다.
- 이렇게 구성된 각 발현 데이터 셀에 대해 1단계에서 생성한 여러 유전자-집합을 각각 고려하여 ES를 계산함으로써 ES에 관한 귀무분포(null distribution)를 생성한다.
- 이러한 귀무분포를 이용하여 1단계에서 계산한 ES의 nominal P-value를 계산한다.

[3 단계] 다중 검증(multiple hypothesis test)에 대한 통계적 유의 수준을 조정하는 단계

- 전체적인 분석 과정에서 여러 유전자-집합에 대한 각각의 유의성 판단이 필요하므로 다중 검증을 위해 2단계에서 추정된 유의 수준을 조정할 필요가 있다.
- 먼저, 각 유전자-집합에 대하여, 이전 단계에서 구한 ES를 유전자-집합의 크기에 대해 정규화함으로써 NES(normalized ES)를 생성한다.
- 각 NES에 대해 False Discovery Rate (FDR)을 계산함으로써 false positive의 비율을 조절하고[1], 이에 따라 통계적으로 유의한 유전자-집합을 최종 결정한다.

III. Fisher Criterion 기반 유전자 서열화 기법을 적용한 GSEA 분석

GSEA 분석을 통해 두 클래스 간에 차이를 나타내는 유의한 유전자-집합을 찾아내기 위해서는, 앞에서 기술한 바와 같이, 유전자 주석 데이터베이스를 기반으로 특정 카테고리에 속한 유전자들로 구성된 다양한 유전자-집합을 생성한 후 각 유전자-집합에 대한 ES 값을 계산할 필요가 있다. 이 과정에서 실험에 사용된 전체 유전자에 대해 두 클래스 간의 발현값 차이에 따라 유전자 서열화 기법을 적용하여 전체 유전자를 정렬하고, 정렬된 전체 유전자 리스트와 각 유전자-집합과의 Kolmogorov-Smirnov Score[2]를 계산함으로써 각 유전자-집합의 ES를 구한다[1~3]. 이 때 전체 유전자를 정렬하기 위한 유전자 서열화 방법으로 식 (1)과 같은 SNR 기법을 주로 사용하고 있다.

$$SNR(i) = \frac{\mu_A(i) - \mu_B(i)}{\sigma_A(i) + \sigma_B(i)} \quad (1)$$

위의 식(1)은 전체 유전자 리스트에 속한 특정 유전자 i에 대하여 서로 다른 두 클래스 A와 B에 속한 샘플들의 발현값 평균을 각각 μ_A와 μ_B라 하고 표준편차를 σ_A와 σ_B라 할 때, 특정 유전자 i의 SNR값을 계산하는 식을 나타낸 것이다. 이러한 SNR을 이용하여 유전자를 서열화할 경우, 그림 1과 같이 정렬되는 특징이 있다.

즉, 두 클래스 A, B 중 클래스 A에 속하는 샘플에서

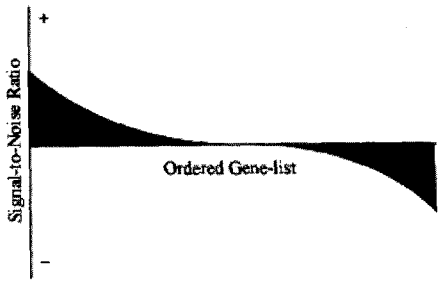


그림 1. Signal-to-Noise Ratio에 의한 유전자 서열화 적용 결과의 예

Fig. 1. Example of Gene Ranking by Signal-to-Noise Ratio.

1) 이 때, FDR은 주어진 NES를 가지는 유전자-집합이 false positive일 확률을 추정한 것으로써, NES의 귀무분포와 관찰된 NES 값의 tail을 비교함으로써 계산된다.

2) Kolmogorov-Smirnov Score는 정렬된 전체 유전자 리스트와 특정 유전자-집합의 분포가 얼마나 유사한가를 나타내는 척도이다.

의 유전자 발현값이 다른 클래스 B에 비해 높을 경우 SNR이 양의 값을 가지며, 반대로 낮을 경우 SNR값은 음의 값을 가지게 된다. 그리하여, SNR 값이 양의 영역(positive region)에서는 값이 클수록, 음의 영역(negative region)에서는 값이 작을수록 두 클래스 간의 발현 차이가 높은 것을 의미한다. 이러한 특징을 지닌 SNR을 기반으로 유전자 서열화를 수행한 뒤 ES 값을 계산하여 유의 유전자-집합을 검출할 경우, 각 유전자-집합에서는 ES 값이 0으로부터 최대의 편차(deviation)를 가지는 점을 기준으로 그 기준점에 따라 양의 영역이나 음의 영역 중의 어느 한 영역에 속하는 유전자들을 유의 유전자들로서 고려하게 된다. 따라서 각 유전자-집합 내에서의 유의 유전자들은 발현값이 클래스 A에서 상대적으로 높게 나타나는 유전자들만의 그룹이거나 혹은 낮게 나타나는 유전자들만의 그룹으로 이루어지기 때문에, 이 두 가지 특성을 지닌 유전자들이 혼용되어 동시에 선택되어지지 못하는 문제가 있다.

그리하여, 본 논문에서는 어느 한 클래스에서의 유전자 발현값이 다른 클래스에 비해 많고 적음보다는 두 클래스 간에 많은 발현값의 차이를 나타내는 유전자들을 유의 유전자로서 고려될 수 있도록 아래 식 (2)와 같은 Fisher criterion[4~6]에 기반한 유전자 서열화 기법을 적용하였다. 이를 통해, 클래스 A의 발현값이 B에 비해 높고 낮음에 상관없이 두 클래스 간에 차이가 많이 나는 유전자들이 유의 유전자로서 고려되고, 각 유전자-집합의 유의성을 판단하는 데에 반영되도록 하였다. 특정 유전자 i에 대한 *Fisher Criterion* 계산식은 다음과 같다.

$$Fisher\ Criterion(i) = \frac{(\mu_A(i) - \mu_B(i))^2}{\sigma_A(i)^2 + \sigma_B(i)^2} \qquad (2)$$

위와 같은 Fisher criterion을 이용하여 유전자를 서열화할 경우, 그림 2와 같이 정렬되는 특징을 보여준다.

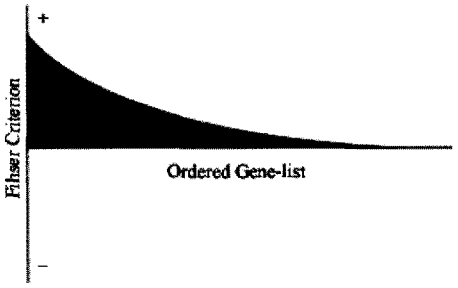


그림 2. Fisher Criterion에 의한 유전자 서열화 적용 결과의 예

Fig. 2. Example of Gene Ranking by Fisher Criterion.

Fisher Criterion 기반 유전자 서열화 기법에 의해 정렬된 전체 유전자 리스트를 이용하여, 각 유전자-집합별로 ES를 계산한다. *ES*를 구하기 위해서는 아래 식 (3)과 같이 전체 유전자 리스트 중 i번째 유전자에 관한 $P_{hit}(i)$와 $P_{miss}(i)$를 계산하고, 전체 유전자 중에서 이 두 값의 절대값 차이가 가장 클 때를 현재 유전자-집합의 *ES* 값으로 정의하였다.

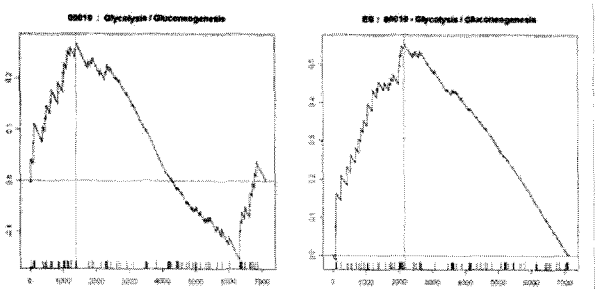
$$P_{hit}(i) = \sum_{j=1}^{i} \frac{E(j)}{N_H}$$

$$P_{miss}(i) = \sum_{j=1}^{i} \frac{(1 - E(j))}{N_M}$$

$$ES = \max_{i=1,\cdots,N} |P_{hit}(i) - P_{miss}(i)| \qquad (3)$$

여기서 $P_{hit}(i)$는 전체 유전자 리스트에서 i번째 유전자가 유전자-집합에 포함되었을 때 사용하는 식이고, 반대로 $P_{miss}(i)$는 i번째 유전자가 유전자-집합에 포함되지 않았을 때 사용하는 식이다. $E(j)$는 해당 유전자-집합 내에서 j번째에 해당하는 유전자의 Fisher Criterion이다. N_H는 해당 유전자-집합에 속한 유전자의 개수이고, N_M는 전체 유전자 리스트에서 해당 유전자-집합에 속하지 않는 유전자의 개수이다. 한편, N은 전체 유전자 리스트에 속한 유전자의 개수이다.

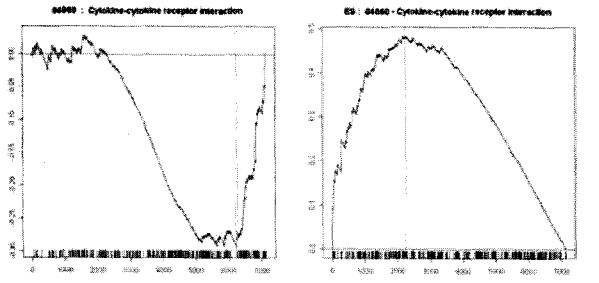
SNR을 이용한 유전자 서열화 기법과 Fisher criterion을 이용한 유전자 서열화 기법은 그림 3, 4, 5에서 나타난 바와 같이 계산된 ES 값에서 차이를 나타낸다. 그림 3은 SNR과 Fisher Criterion의 두 가지 경우에서 모두 ES값이 양의 영역에 존재하는 경우를 나타내며, 그림 4는 SNR의 경우 음의 영역에서 ES값이 선택되고 Fisher criterion의 경우 양의 영역에서 ES값이 선택되는 경우를 보이고 있다. 이처럼 그림 3과 4의 경우에는 모두 0과의 편차(deviation)가 가장 큰 지점을 기준으로 ES를 결정하게 된다. 한편, 그림 5와 같은 경우에는 SNR에서는 문제가 되지 않지만 Fisher criterion의 경우 0과의 편차가 가장 큰 부분이 음의 영역에 존재한다. 그러나 실제 Fisher criterion을 사용할 경우 그 측정치가 항상 양의 값만을 갖기 때문에 양의 영역에서의 분포의 유사성만 판단하면 된다. 따라서 음의 영역에서 0과의 편차가 가장 큰 부분을 제외하고 양의 영역에서 0과의 편차가 가장 큰 부분이 그 유전자-집합의 ES값이 되도록 조정하였다.



SNR Fisher Criterion

그림 3. SNR 기반 유전자 서열화 방법 사용시 ES값이 양의 영역에 존재하는 경우의 예와 Fisher Criterion 적용결과의 비교

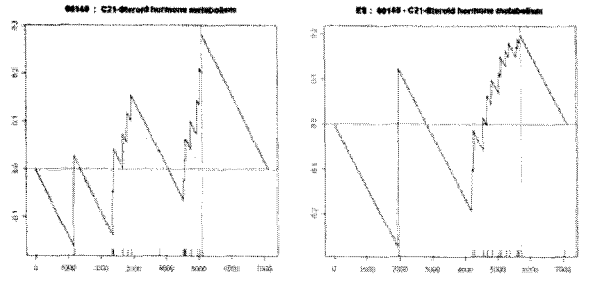
Fig. 3. An example of ES existing in a positive region when SNR-based gene ranking is used, and its comparison to Fisher Criterion's case.



SNR Fisher Criterion

그림 4. SNR 기반 유전자 서열화 방법 사용시 ES값이 음의 영역에 존재하는 경우의 예와 Fisher Criterion 적용결과의 비교

Fig. 4. An example of ES existing in a negative region when SNR-based gene ranking is used, and its comparison to Fisher Criterion's case.



SNR Fisher Criterion

그림 5. Fisher Criterion 기반 유전자 서열화 기법 이용시 ES값이 음의 영역에 존재하는 경우의 예와 SNR 적용결과의 비교

Fig. 5. An example of ES existing in a negative region when Fisher Criterion-based gene ranking is used, and its comparison to SNR's case.

Ⅳ. 실 험

본 논문에서는 상기에서 제안한 Fisher criterion 기반 GSEA를 통한 유의 유전자-집합 및 유전자를 검출하는 실험을 위해 1999년 Golub et al.[15]에 의해 발표된 바 있는 Leukemia 데이터 셑[5]을 사용하였다. 이 데이터 셑은 백혈병의 두 가지 다른 클래스인 Acute Myeloid Leukemia(AML) 및 Acute Lymphoblastic Leukemia(ALL)에서 나타나는 인간 유전자 7129개의 발현 프로파일로 구성되어 있다. 총 38개의 실험 샘플이 사용되었으며, 27개가 ALL 클래스에 해당하고 나머지 11개는 AML 클래스에 해당한다.

본 실험에서는 DNA 마이크로어레이 실험을 통해 획득한 Leukemia 관련 유전자 발현 프로파일을 이용하여 AML과 ALL을 구분할 수 있는 유의한 유전자-집합 및 유전자들을 검출하기 위해, 유전자 서열화 방법으로 SNR을 사용한 GSEA와 Fisher criterion을 사용한 GSEA를 비교 분석하고자 하였다. 이를 위하여, 먼저 KEGG Pathway[16~18]를 이용하여 167개의 유전자-집합을 구성하였고, 이 중 AML과 ALL 두 클래스에서 중요한 발현값의 차이를 나타내는 유의한 유전자-집합을 찾아내고 이들의 패스웨이 기능을 분석하였다. 또한, 이렇게 검출된 유의 유전자-집합의 패스웨이 기능이 기존에 알려진 Leukemia와 관련된 패스웨이인지를 비교 분석하였다.

4.1 KEGG Pathway 기반 유전자-집합의 구성

Leukemia 데이터 셑에 속한 전체 유전자 리스트 중에서 유전자 주석 데이터베이스에 의해 기능이나 관계를 나타내는 카테고리를 이용하여 유전자를 분류한 것이 유전자-집합이다. 실험에서 사용한 유전자 주석 데이터베이스는 KEGG pathway를 이용하였다. 유전자-집합을 구성하기 위해 KEGG pathway 데이터베이스에서 pathway 기능별로 카테고리화한 데이터들을 이용하여 특정 pathway에 관련된 유전자들로 유전자-집합을 구성하였다[19~20]. 유전자-집합을 구성할 때 유전자의 수가 너무 작은 것은 GSEA 분석에서 잘못된 결과를 얻을 수 있기 때문에 pathway 기능별로 분류된 유전자-집합을 구성하는 유전자의 수가 최소 5개 이상이 되는 167개의 유전자-집합들을 구성하여 GSEA분석에 이용하였다.

4.2 Fisher Criterion 기반 유전자 서열화 기법에 의한 유의 유전자 선정 결과 분석

상기에서 제안된 Fisher Criterion 방법에 의해 유전자 서열화를 수행할 경우 동일한 유전자-집합에 대해서 기존의 SNR 방법과는 다른 유의 유전자를 선정하게 된다. 아래 그림 6은 Fisher Criterion에 의한 유전자 서열화 기법과 SNR 기반 유전자 서열화 기법을 GSEA 분석에 사용할 경우, KEGG Pathway ID가 has04662인 유전자-집합 "B cell recepter signaling pathway"에서 선정된 유의 유전자 결과를 비교한 것이다. SNR 측정치를 기준으로 할 때, 그림 6 (a)에서는 위쪽에 위치한 유전자일수록 두 클래스 간에 발현값 차이가 큰 것을 나타내며, 그림 6(b)에서는 아래쪽에 위치할수록 발현값 차이가 큰 유전자를 나타낸다. 특히, 그림 6에서 밝은 회색으로 표기된 유전자들은 유전자-집합 "B cell recepter signaling pathway"에 대해 SNR을 이용한 GSEA 분석시 유의 유전자로 선정된 결과를 나타내며,

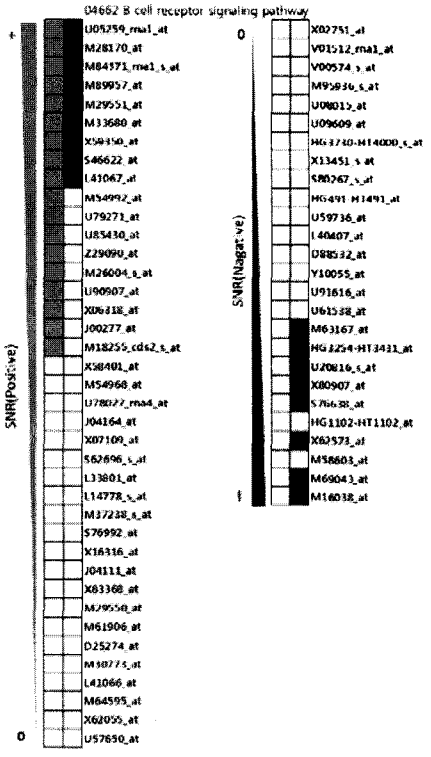


그림 6. SNR vs. Fisher Criterion 기반 GSEA 분석에 의한 유의 유전자 선정 결과 비교 (어두운 회색: Fisher Criterion 결과, 밝은 회색: SNR 결과)

Fig. 6. A comparison of statistically significant genes chosen by SNR vs. Fisher Criterion based GSEA.(dark grey boxes: Fisher Criterion results, white grey boxes: SNR results).

어두운 회색으로 표기된 유전자들은 Fisher Criterion 기반 GSEA 적용시 유의한 유전자들로 최종 선정된 결과를 나타낸 것이다.

상기 그림 6에서와 같이, GSEA 분석을 위해 SNR을 유전자 서열화 기법으로 적용할 경우에는 두 클래스 중 어느 한쪽의 발현값이 많이 나타나거나 적게 나타나는, 즉 SNR 값이 음의 영역이나 양의 영역 중에 어느 한 영역에 속하는 유전자들만을 유의 유전자로 고려하는 특징이 있다. 반면에, Fisher Criterion의 경우 측정치의 특성상 두 클래스에서의 발현값이 어느 한 쪽에서 많고 적음보다는 두 클래스 간의 발현값 차이가 크고 작음에 따라 유의 유전자를 결정하고 이에 기반하여 유전자-집합의 유의성을 판별하는 특징을 지닌다. 따라서 Fisher criterion 기반 유전자 서열화 기법을 GSEA 분석에 사용할 경우, SNR 사용 시에 놓칠 수 있는 중요한 유의 유전자-집합 검출에 도움이 될 것으로 추정된다.

4.3 유의 유전자-집합의 선정 결과 및 생물학적 의미 해석

일반적인 GSEA 분석에서 유의한 유전자-집합을 찾는 방법에는 정규화된(normalized) ES, 즉 NES값을 이용하여 p-value를 유의수준으로 판단하여 유의한 유전자-집합들을 추출하는 방법과, Family-wise Error Rate(FWER)와 False Discovery Rate(FDR) 등과 같은 다중 검증을 이용하는 방법이 있다. 본 실험에서는 NES값을 이용하여 상위 40 개에 해당하는 유전자-집합들을 유의한 유전자-집합으로 선정하고, 백혈병과 관련하여 이미 알려진 KEGG 패스웨이 정보와 GSEA분석에 의해 검출된 유의 유전자-집합들이 상호 얼마나 매칭 되는지를 검토하여 기존 SNR 기반 GSEA와 새로운 Fisher Criterion 기반 GSEA 결과들을 비교 평가하였다.

특히 KEGG 패스웨이에서 백혈병의 AML 클래스와 관련이 있다고 밝혀진 패스웨이 *hsa04110*과 *hsa04210* (hsa04110: Cell cycle, hsa04210: Apotosis), ALL 클래스와 관련된 *hsa04660*와 *hsa04662* (hsa04660: T cell receptor signaling pathway, hsa04662: B cell receptor signaling pathway), 그리고 AML과 ALL 클래스 둘 다 동시에 관련이 있다고 밝혀진 *hsa04640* (hsa04640: Hematopoietic cell lineage)[21-22]을 선정하여 백혈병 관련 패스웨이로 고려하였다. 이를 기반으로 각 GSEA 방법들을 분석을 통해 유의성이 높다고 추출된 상위 40개의 유전자-집합 중에서, 백혈병 관련 다섯 개의 패스웨

표 1. Fisher Criterion 기반 유전자 서열화 기법과 SNR기반 유전자 서열화 기법을 적용한 GSEA 실험 결과

Table 1. GSEA results obtained by Fisher Criterion and SNR[1] based gene ranking methods.

유전자 서열화 기법	Type of Leukemia	Identified KEGG Pathway ID.
SNR 기반 GSEA	AML	hsa04110
	ALL	
	ALL & AML	hsa04640
Fisher Criterion 기반 GSEA	AML	hsa04110 hsa04210
	ALL	hsa04660 hsa04662
	ALL & AML	hsa04640

이를 얼마나 포함하고 있는지를 검토하여 결과를 검증하였다. 이에 관한 분석 결과는 표 1과 같다.

상기 표 1에 나타난 바와 같이, 기존의 SNR을 GSEA 분석에 이용할 경우, AML 관련 패스웨이 정보로서 hsa04110와 ALL/AML 관련 패스웨이로서 hsa04640인 총 두 개의 패스웨이만을 추출하였다. 반면에, 새로운 Fisher Criterion 을 이용한 경우에는 AML 관련해서는 hsa04110과 hsa04210, ALL과 관련된 hsa04660과 hsa04662, 그리고 ALL/AML에 관련된 hsa04640 패스웨이를 추출함으로써 생물학적으로 이미 검증된 총 다섯 개의 백혈병 관련 패스웨이 정보를 모두 추출하였다.

V. 결 론

이상에서와 같이 본 논문에서는 GSEA 분석을 위한 유전자 서열화 방법으로서 일반적으로 사용되는 SNR 방법 대신 Fisher Criterion을 이용하는 방법을 제안하고, 이의 유용성을 백혈병 관련 실험 분석을 통해 살펴보았다. 실험 결과에 의하면 SNR 기반 유전자 서열화 방법을 이용한 경우 백혈병의 두 클래스인 AML과 ALL에서 중요한 발현값의 차이를 보이는 유의한 패스웨이로서 실제 생물학적으로 관련성이 밝혀진 총 5개의 패스웨이 중 2개만을 찾아낼 수 있었다. 반면에 Fisher Criterion을 사용한 경우에는 5개 모두를 찾아낼 수 있었다. 따라서 특정 클래스에서의 유전자 발현값이 다른 클래스에 비해 상대적으로 높거나 낮은 패턴을 동시에 반영하는 Fisher Criterion을 GSEA 분석에 적용할 경우 SNR을 이용할 때에 발견하지 못한 생물학적으로 의미 있는 유의 유전자-집합을 효과적으로 추출할 수 있었다.

참 고 문 헌

[1] A. Subramanian et al., "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. Taskesen, "Sub-typing of model organisms based on gene expression data." Bioinformatics technical University of Delft Research Assignment, 2006.

[3] S. Monti et al., "Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response.", Blood. 2005 Mar 1;105(5):1851-61, Nov 2004.

[4] C. Bishop, "Neural Networks for Pattern Recognition", Oxford University Press, Oxford, 1995.

[5] A. Blum et al., "Selection of relevant features and example in machine learning", Artificial intelligence, 97:245-271, 1997.

[6] P. Bradley et al., "Feature selection via mathematical programming", Technical report to appear in INFORMS Journal on computing, 1998.

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

[8] S. Dudoit et al., "Multiple Testing Procedures and Applications to Genomics", Springer, 2007.

[9] G. J. McLachlan et al., "ANALYZING MICROARRAY GENE EXPRESSION DATA", WILEY-INTERSCIENCE John Wiley & Sons, 2004.

[10] S. Dudoit et al., "Multiple Hypothesis Testing in Microarray Experiments", Statistical Science, 18: 71-103, 2003.

[11] Y. Ge et al., "Resampling-based multiple testing for microarray data analysis", Technical Report 633, Department of Statistics, University of California, Berkeley, 2003.

[12] V. G. Tusher et al., "Significance analysis of microarrays applied to the ionizing radiation response", Proc Natl Acad Sci. 24;98(9):5116-21, Apr 2001.

[13] R. Gentleman et al., "Bioinformatics and Computational Biology Solutions Using R and

Bioconductor", Springer, 2005.
[14] J. Verzani, "Using R for Introductory Statistics" Chapman & Hall/CRC, Boca Raton, FL, 2005.
[15] T. R. Golub et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science (Wash. DC), 286: 531.537, 1999.
[16] KEGG: Kyoto Encyclopedia of Genes and Genomes , http://www.genome.ad.jp/kegg/
[17] M. Kanehisa et al., "The KEGG databases at GenomeNet, Nucleic Acids Res.", 30:42-46, 2002.
[18] S. Kawashima et al., "KEGG API: A Web Service Using SOAP/WSDL to Access the KEGG System", Genome Informatics 14: 673-674, 2003.
[19] I. Dinu et al., "Improving GSEA for analysis of biologic pathways for differential gene expression across a binary phenotype.", Collection of Biostatistics, 2007.
[20] T. Manoli et al., "Group testing for Pathway analysis improves comparability of different microarray datasets", Bioinformatics, 22(20):2500-2506, 2006.
[21] S. Kudsen, "Cancer Diagnostics with DNA Microarrays", John Wiley & Sons, Inc., 2006.
[22] C. Potten et al., "Apoptosis", Cambridge University Press, 2005.

저 자 소 개

김 재 영(학생회원)
2006년 위덕대학교 컴퓨터공학과 학사 졸업.
2008년 경북대학교 대학원 정보통신학과 석사 수료
<주관심분야 : 생물정보학, 데이터마이닝, 패턴인식>

신 미 영(정회원)
1991년 연세대학교 전산과학과 학사 졸업.
1993년 연세대학교 전산과학과 석사 졸업.
1998년 미국 Syracuse Univ., EECS Dept., Ph.D
1999년~2005년 3월 한국전자통신연구원 선임 연구원
2005년 4월~현재 경북대학교 전자전기컴퓨터학부 조교수
<주관심분야 : 패턴인식, 바이오인포매틱스>