ISSN 1598-9798

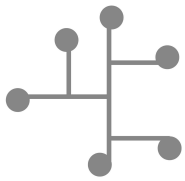
데이타베이스연구
제28권 제3호 2012년 12월

유전자 발현량 차이를 이용한 네트워크 기반 질병 관련 유전자 탐색 기법

A Network-based Approach to Detect Disease-related Genes using Differentially Expressed Gene Analysis

김현진, 안재균, 박상현
Hyunjin Kim, Jaegyoon Ahn, Sanghyun Park

데이터베이스 소사이어티
Database Society
사단법인 한국정보과학회
The Korean Institute of Information Scientists and Engineers



유전자 발현량 차이를 이용한 네트워크 기반 질병 관련 유전자 탐색 기법

A Network-based Approach to Detect Disease-related Genes using Differentially Expressed Gene Analysis

김현진(Hyunjin Kim)[1], 안재균(Jaegyoon Ahn)[2], 박상현(Sanghyun Park)[3]

요 약

마이크로어레이 데이터를 분석하는 대표적인 방법 중 한가지는 차등 발현 유전자(Differentially Expressed Gene)들을 찾는 것이다. 차등 발현 유전자란 두 실험 조건 하에서 샘플 집합의 유전자 발현량이 많이 차이나는 유전자를 의미한다. 하지만 기존의 차등 발현 유전자를 찾는 방법들은 유전자끼리 주고 받는 영향을 고려하지 않아 근본적인 한계를 지니고 있다. 실제로 기존 방법들로, 질병과 관련되어 있다고 생물학적 실험으로 증명된 유전자들을 많이 찾아내지 못하고 있다. 이러한 한계를 극복하기 위해 본 연구에서는 유전자 사이의 상관관계를 고려한 단백질-단백질 상호작용(Protein-Protein Interaction) 네트워크를 적용하여 유전자 간에 미치는 영향을 기존 방법에 추가함으로써 차등 발현 유전자를 검색하는 새로운 방법을 제안한다. 이렇게 찾아낸 유전자들로 질병과 관련된 클래스 분류를 시도한 결과, 기존의 네트워크적 접근 방법을 적용하지 않은 차등 발현 유전자를 찾는 방법보다 더 높은 정확도와 AUC(Area Under Curve)를 보였다. 또한 점수 값의 상위에 위치해있는 유전자들이 해당 질병과 얼마나 관련되어 있는지에 대해서 다른 특성 선택(Feature Selection) 방법들과 비교해보았을 때 더 낮은 p-value를 나타냄으로써 본 연구의 방법이 질병 관련 유전자를 잘 검색한다는 사실을 보여주었다.

주제어: 특성 선택, 차등 발현 유전자, 마이크로어레이 데이터 분석, 질병, 생물정보학

1 연세대학교 공과대학 컴퓨터과학과, 통합과정
2 연세대학교 공과대학 컴퓨터과학과, 박사과정
3 연세대학교 공과대학 컴퓨터과학과, 교수
† 본 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 중견연구자지원사업(도약연구) 지원을 받아 수행된 것임 (2012-010775).

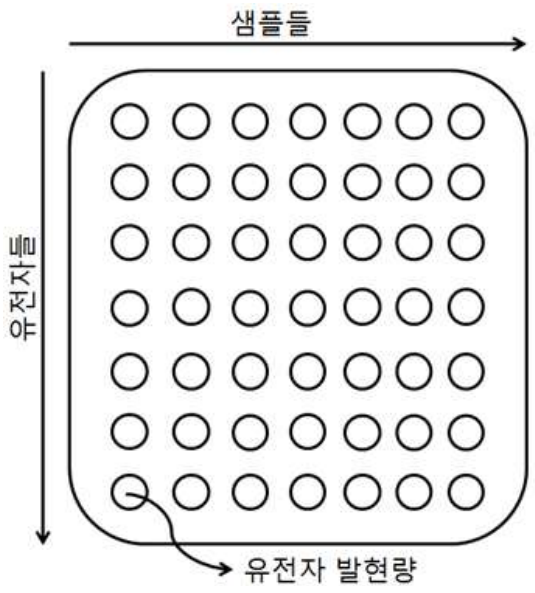
Abstract

One of general method for microarray analysis is discovering differentially expressed genes. The differentially expressed gene is a gene which shows different expression levels between two conditions. However, existing methods for finding differentially expressed genes have a limitation. They cannot consider influences among genes. Specifically, they can hardly discover the biologically proved genes which are related to specific diseases. To get over the limitation, we propose a novel approach to discover disease-related genes using differentially expressed genes and protein-protein interaction network. The proposed approach uses protein-protein interaction to reflect the influences among genes. The approach showed better accuracy and AUC(Area Under Curve) value than a method which does not consider the influences among genes and showed lower p-value than other feature selection methods.

Keywords: Feature selection, Differentially expressed gene, Microarray analysis, Disease, Bioinformatics

1. 서 론

유전자의 발현은 세포 내의 mRNA의 양을 측정하는 것이며, 이것을 위한 가장 유명하고 많이 사용되어 온 실험은 마이크로어레이(Microarray) 기술을 통한 것이다. 마이크로어레이는 유리판 위에 집적되어 있는 유전자들을 대상으로 하여 모든 유전자 발현량을 한번의 실험으로 알아낼 수 있다(그림 1). 마이크로어레이 데이터에서 한 행(Row)은 한 샘플에서 각 유전자들의 발현량을 나타내고, 한 열(Column)은 각 샘플들의 특정 유전자에서의 발현량을 보여준다. 마이크로어레이는 기존 유전자 발현량을 알아내기 위한 실험들보다 양적에서나 분석 속도 측면에서나 더 나은 효율성을 지니고 있어 널리 사용되고 있다.



<그림 1> 마이크로어레이 데이터

마이크로어레이 데이터를 분석하는 방법은 바이오인포매틱스에서 가장 많이 연구되어 온 주제 중 하나인데, 이 가운데서도 가장 기본적이면서도 최근까지도 많이 사용되는 것은 두 실험 조건하에서의 샘플 집합의 유전자 발현량을 비교하는 것이다. 예를 들어, 이러한 분석 방법은 세포 간의 차이를 밝힘으로써 [1] 질병과 관련된 유전자를 찾아내는 수많은 연구에 사용되고 있으며, 세포 발달의 차이를 찾아내는 등의 연구 [2] 나 microRNA의 타겟 유전자를 찾아내는 연구 [3] 에도 적용되고 있다.

하지만 이러한 마이크로어레이 유전자 발현량 데이터에 대한 기존 차등 발현 유전자(Differentially Expressed Gene) 탐색 기술은 근본적인 한계를 가진다. 그것은 유전자의 절대적인 양이나 두 조건하에서 양의 차이에 그 의미를 부여하는 절대적인 기준이 모두 다를 수 있다는 것이다. 예를 들어, 다른 유전자에 비해서 직접적으로 영향을 주고 받는 유전자가 많은 유전자 A가 두 조건에서 큰 발현량 차이를 보이지 못하나 실제로 두 조건의 차이를 잘 설명해 줄 수 있는 유전자라고 할 때 다음과 같은 이유에서 A가 차등 발현 유전자로 선정되지 못할 수 있다. 첫 번째로 A는 조그만 변화로도 많은 유전자에 영향을 미칠 수 있다. 두 번째로, A와 직접적으로 영향을 주고 받는 유전자들의 영향에 의해서 그 발현량의 차이가 측정되지 못한다. RNA-seq 기술을 이용한 유전자 발현량 데이터는 그 노이즈가 줄어들기 때문에 더 정확하게 발현량을 측정할 수 있다는 장점에도 불구하고 마이크로어레이 데이터가 가진 이와 같은 한계점을 극복하지는 못한다.

이러한 한계점을 극복하기 위해서 본 연구에서는 유전자 사이의 상관 관계를 함축하는 유전자 간 네트워크 데이터를 추가적으로 사용하고자 한다. 이러한 유전자 네트워크로써 가장 많이 사용되는 것은 단백질-단백질 상호작용 네트워크(Protein-protein Interaction Network, 이하

PIN)이다. PIN은 서로 간에 물리적으로 결합하는 단백질들을 측정해서 네트워크화 한 데이터이다. PIN은 질병에 특이적인 유전자 및 그 관계를 탐색하는 기존의 연구들 [4, 5, 6, 7, 8] 에도 많이 사용된 바 있는데, PIN을 기반으로 한 마이크로어레이 데이터의 분석은 마이크로어레이 데이터 하나만 사용한 경우보다 훨씬 높은 정확도로 샘플을 분류하며 질병과 관련된 유전자를 더욱 잘 찾음을 보여준 바 있다. 본 연구에서는 기존의 차등 발현 유전자 탐색 방법을 이용해서 각 유전자가 두 조건의 차이를 설명해 줄 수 있는 정도를 우선적으로 스코어링 한 후, 네트워크 전달 (Network Propagation) 방법을 통해서 각 유전자들이 네트워크 상으로 연결된 다른 유전자들에게 미치는 영향을 반영하여 효율적으로 질병과 관련된 유전자들과 차등 발현 유전자들을 검색하고자 한다.

2. 관련 연구

마이크로어레이 데이터에서 두 조건하에서의 유전자 발현량을 비교할 때는 그 발현량이 비교적 큰 차이를 보이는 유전자를 많이 연구하게 되므로, 이를 찾아내는 절차가 필수적이다. 이러한 과정을 차등 발현 유전자 탐색이라고 한다. 전통적으로 이러한 탐색 방법은 크게 두 가지로 나뉜다. 하나는 발현량의 배수 차이 (Fold Change)를 측정하는 방법이며, 한 가지는 t-검정을 이용한 p-value를 구하는 것이다 [9].

유전자 발현량의 배수 차이는, 단순히 두 조건 집합 내의 샘플들의 유전자 발현량의 평균이 일정 배수 이상의 차이를 보이는 경우를 의미한다 [10]. 그러나 이러한 방식은 마이크로어레이의 플랫폼이나, 실험 조건 등 많은 요인에 의해서 유전자 발현 데이터에 따라서 많은 편차를 보인다. 이러한 편차를 제거해 주기 위해서 많이 시도된 방법은 오프셋(Offset)을 이용한 배수 차이 방법인데 [11, 12, 13], 이 방법은 배수 비교를 계산하기 전에 일정 오프셋을 더하는 방법을 의미한다.

p-value를 구하는 방식은 t-검정을 이용해서 실험(Test) 및 비교(Control) 집합이 차이를 보이지 않는다는 귀무 가설을 채택할 확률을 구하는 방식이다. 그러나, 이러한 방식은 유전자가 두 조건에서 얼마나 차이를 보이는가를 말해주지는 않는다. 또한, 샘플의 개수가 많을 경우 조건 간의 미미한 차이에서도 통계적으로 유의미한 p-value를 보일 수 있다 [14].

배수 차이와 p-value 방식은 때로 합쳐져서 사용되기도 하는데, 이러한 대표적인 방식에는 볼카노 플롯 (volcano plot) [15] 등이 있다. 그러나 이러한 방법은 유의미한 유전자 순위를 제공해주지는 않는다는 한계를 가진다.

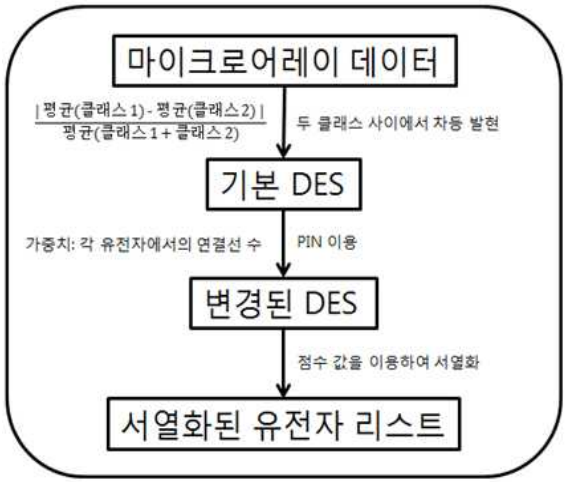
최근에는 RNA-seq 기술이 기존의 마이크로어레이 기술을 대체하고 있다. RNA-seq는 세포 내의 mRNA를 High-Throughput 시퀀싱(Sequencing)을 사용해서 측정하는 방식으로써, 시퀀싱 비용의 감소에 따라서 유전자 발현량 분석에서 더욱 빈번하게 사용되어 가고 있다 [16, 17, 18, 19]. RNA-seq 방식의 유전자 발현량 데이터는 기존 마이크로어레이 데이터에 비해서 노이즈 등의 기술적 편차가 감소되었다 [20]. 하지만 마이크로어레이 데이터가 복제된 여러 개의 샘플에 대한 유전자 발현량을 측정해서 통계적인 유의미성을 보장하는 반면, RNA-seq을 이용한 유전자 발현량 데이터의 70% 가까이가 단 하나의 샘플에 대한 유전자 발현량을 측정한 것이다. 이러한 문제에 대

해서 Feng 등 [9] 은 각 샘플에 기존의 로그 배수 (Log Fold) 변화의 분포에 의거해서 안정적인 통계량을 부여하는 차등 발현 유전자 탐색 방식인 GFOLD를 제안했다. 또한, RNA-seq 방식은 리드가 다수의 유전자 혹은 엑손 (Exon)에 매핑된다는 불확실성이 존재하는데, Glaus 등 [21]은 기존의 데이터를 이용해서 생물학적 편차를 모델링함으로써 이러한 불확실성을 제거하는 차등 발현 유전자 탐색 방법을 제안하였다.

3. PIN 기반 질병 관련 유전자 탐색 기법

본 연구에서는 기존의 차등 발현 유전자에 네트워크적 접근을 이용하여 질병 관련 유전자를 더 효율적으로 찾을 수 있는 방법을 제안한다. 각 유전자들에서 기본적인 차등 발현 점수(Differentially Expressed Score, 이하 DES)를 구하고 PIN을 이용하여 각 유전자에 연결된 모든 점수를 정규화하고 더하면 우리가 원하는 새로운 DES를 얻을 수 있다. 또한 새로운 DES를 이용하여 유전자들의 서열을 매길 수 있다 (그림 2).

기본적인 DES를 구할 때는 서로 다른 두 실험 조건상의 샘플 집합들의 평균을 구하고 두 평균간의 차이를 이용한다. 유전자 사이의 발현량 차이로 인해 기본적으로 발현량이 높게 나오는 유전자의 경우, 실질적으로 두 조건상의 차이가 많이 나지 않는 것이라 해도 다른 기본적 발현량이 낮게 나오는 유전자들보다 더 차등적으로 발현된다고 여겨질 수 있으므로 해당 유전자의 전체 평균으로 그 차이를 나누어 정규화한다. 변경된 DES는 PIN상에서 연결된 모든 유전자들의 DES를 기존 DES에 합산하여 계산한다. 이 때, 연결선이 많은 유전자의 경우 변경된 DES가 잘 나올 수 있으므로 기본적인 DES에 해당 유전자에 연결된 다른 유전자들의 DES를 더하기 전에 가중치로 (1 / 해당 유전자의 연결선 수)를 곱해준다. 본 연구의 방법을 알고리즘으로 나타낼 수 있다 (알고리즘 1).



<그림 2> 본 연구 방법의 흐름

3.1 기본 DES

본 연구의 방법은 네트워크적 접근을 통해서 유전자의 상호간 영향을 고려하고자 하는 목적을 가지고 있다. 따라서 각 유전자들의 기본적인 점수와 네트워크를 통해 영향을 미치게 될 점수를 결정해야 한다. 기본적인 점수 의 경우, 각 클래스 별로 유전자 발현량을 모두 더한 후, 평균을 내서 그 평균의 차이로 결정한다. 하지만 이때 단순히 평균의 차이를 이용한다면 유전자간의 발현량 차이를 고려하지 않게 된다.

예를 들어 유전자 A의 경우 클래스 1의 유전자 발현량이 450, 500, 550이고 클래스 2의 유전자 발현량이 350, 400, 450 이며, 유전자 B의 경우 클래스 1의 유전자 발현량이 34, 35, 36 이고 클래스 2의 유전자 발현량이 2, 3, 4 라면, 유전자 A의

입력 : *N* 개의 유전자로 이루어진 마이크로어레이 데이터 *M*, PPIN, 특성의 개수 *K*
출력 : 서열화된 *K* 개 특성

1. **For each** 유전자 *G(i)* **Do**
2. **For each** 클래스 1의 샘플들 **Do**
3. 클래스 1의 *G(i)* 평균 = 클래스 1 샘플들의 *G(i)* 발현량의 합 / 클래스 1 샘플들의 수
4. **End For**
5. **For each** 클래스 2의 샘플들 **Do**
6. 클래스 2의 *G(i)* 평균 = 클래스 2 샘플들의 *G(i)* 발현량의 합 / 클래스 2 샘플들의 수
7. **End For**
8. *G(i)*의 기본 DES = (|클래스 1의 *G(i)* 평균 - 클래스 2의 *G(i)* 평균|) / *G(i)*의 전체 발현량 평균
9. **End For**
10. **For each** 유전자 *G(i)* **Do**
11. *G(i)*의 변경된 DES = PIN상에서 연결된 모든 유전자들의 기본 DES들의 합 * *G(i)*의 연결선 수
12. **End For**
13. 변경된 DES를 기반으로 유전자들을 서열화
14. 상위 *K*개 특성 선택

<알고리즘 1> 본 연구 방법의 알고리즘

두 클래스 간 평균의 차이는 50이고 유전자 B의 두 클래스 간 평균의 차이는 32이다. 값 차이로 본다면 유전자 A가 더 차등 발현되었다고 볼 수 있지만 실제로 차이가 많이 나는 것은 유전자 B이다.

이러한 상황을 고려하기 위해 단순히 조건 간의 평균 차이가 아니라 평균 차이를 전체 평균으로 나눈 값을 기본 DES로 사용하게 된다. 위의 예제의 경우 유전자 A의 기본 DES는 약 0.22이고 유전자 B의 기본 DES는 1.68로 유전자 B가 더 차등 발현된 유전자임을 알 수 있다. 기본 DES를 구하는 방법을 수식으로 표현하자면 우측 수식과 같다.

이렇게 구해진 기본 DES들은 자신의 점수를 의미함과 동시에 각 유전자들이 다른 유전자들에 미치는 영향력의 크기라고도 볼 수 있다. 따라서 다음 단계에서 네트워크의 연결선을 따라 다른 유전자의 점수에 영향력을 합산하게 된다.

$$des(g_i) = \frac{\sum_{j=1}^{n(C1)} a_j}{n(C1)} \Big/ \frac{\sum_{j=1}^{n(C2)} b_j}{n(C2)}$$

*g_i = i 번째 유전자

* $n(C1)$ = 클래스 1 샘플의 개수

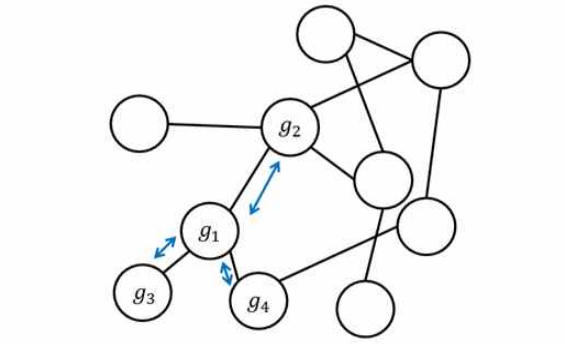
* $n(C2)$ = 클래스 2 샘플의 개수

*a_j = 유전자 g_i의 클래스 1에서 j 번째에 있는 샘플의 발현량

*b_j = 유전자 g_i의 클래스 2에서 j 번째에 있는 샘플의 발현량

3.2 변경된 DES

이전 단계에서 각 유전자들의 기본 점수를 구하였다면 이제 네트워크를 이용하여 유전자들이 서로에게 미치는 영향력을 고려해야 하는 단계이다. PIN을 사용하면 어떤 유전자가 어느 유전자에 영향을 미치는지 알 수 있다. PIN 상에서 특정 유전자에 연결선이 존재한다면 해당 연결선에 연결된 모든 유전자들의 기본 DES 에 특정 유전자의 기본 DES 를 더한다. 이때 정규화를 위해 가중치를 곱해주게 되는데 가중치는 1 / 해당 유전자의 연결선 수이다. 변경된 DES를 구하는 예제가 아래에 표현되어 있다 (그림 3).



<그림 3> PIN에서 변경된 DES를 구하는 예제

3.3 특성 서열화

이전 단계에서 구한 변경된 DES는 유전자들의 영향력을 고려한 점수로써 이를 가지고 유전자들을 서열화 할 수 있다. 유전자는 변경된 DES가 클수록 상위에 위치하게 된다. 상위에 위치한 유전자들은 실질적으로 두 조건을 잘 고려한 유전자들이라고 볼 수 있으며 해당 질병과 관련이 있다고 예측할 수 있다.

4. 결과

본 연구 방법의 실험을 위해 Windows 7 OS 환경의 Intel® Core™ i3 530 Dual 2.93 GHz, 4.00 GB RAM PC를 사용하였고, 방법은 JDK 6 기반의 JAVA언어로 구현하였다. 실험에 사용된 마이크로어레이 데이터는 Singh [22]의 데이터로 8828개의 유전자 발현량을 나타내고 있다. 또한 데이터는 52개의 전립선 암 환자의 샘플과 50개의 일반 환자 샘플로 이루어져 있다. 네트워크 기반 기법을 적용하기 위해 사용한 네트워크는 Brown [23] 의 Protein-Protein Interaction 네트워크를 사용하였다.

본 연구 방법의 검증을 위해 수행한 실험들은 두 가지이다. 첫 번째로, 네트워크 기반 기법을 적용하기 전의 기본 DES로 서열화한 유전자들과 네트워크 기반 기법을 적용한 후의 변경된 DES로 서열화한 유전자들 간의 분류 정확도 및 AUC의 비교이다. 두 번째는, 네트워크 기반 기법을 적용한 방법이 기존 특성 선택 방법들보다 얼마나 질병 관련 유전자들을 잘 찾아내는지에 대한 p-value의 비교이다.

서열화한 유전자들 간의 분류 비교는 SVM(SMO, LibSVM) [24-26], k-nearest-neighbor [27], Random forest [28], Naïve Bayesian [29], 그리고 Bayesian network [30-31] 방법들로 비교하였다 (표 1-2).

분류 비교는 각각의 스코어링 방법으로 상위 500개의 유전자를 선택하여 10-fold Cross Validation으로 수행되었다. 분류 정확도의 경우 SVM(SMO)와 Random forest에서 우수한 성능을 보였으며 AUC는 SVM(SMO), Random forest, Naïve Bayesian, 그리고 Bayesian

network에서 더 높은 값을 나타내었다. 네트워크 기반 기법을 적용하면 분류에 최적화된 기존 차등 발현 방법의 점수와 많이 달라지게 되는데 실험을 수행한 결과, 동일한 결과를 보여주거나 더 우수한 성능을 나타내었다.

<표 1> 기본 DES와 네트워크 기반 기법을 적용한 변경된 DES의 분류 정확도 비교

	기본 DES	네트워크 기반 기법 적용 DES
SVM (SMO)	90.20 %	**95.10 %**
SVM (LibSVM)	83.33 %	83.33 %
k-nearest -neighbor	81.37 %	81.37 %
Random Forest	88.24 %	**91.18 %**
Naïve Bayesian	86.27 %	86.27 %
Bayesian Network	86.27 %	86.27 %

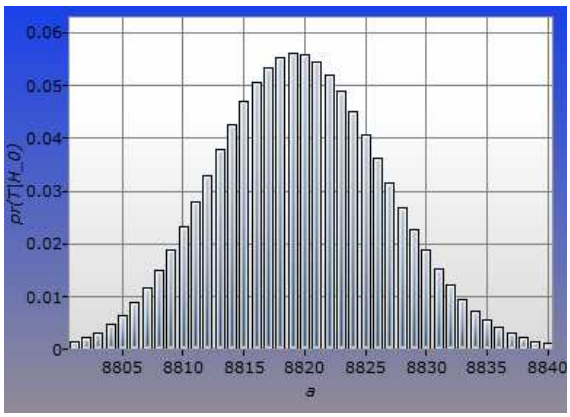
<표 2> 기본 DES와 네트워크 기반 기법을 적용한 변경된 DES의 분류 AUC 비교

	기본 DES	네트워크 기반 기법 적용 DES
SVM (SMO)	0.902	**0.951**
SVM (LibSVM)	0.834	0.834
k-nearest -neighbor	0.824	0.824
Random Forest	0.922	**0.948**
Naïve Bayesian	0.887	**0.893**
Bayesian Network	0.883	**0.894**

분류 비교는 각각의 스코어링 방법으로 상위 500개의 유전자를 선택하여 10-fold Cross Validation으로 수행되었다. 분류 정확도의 경우 SVM(SMO)와 Random forest에서 우수한 성능을 보였으며 AUC는 SVM(SMO), Random forest, Naïve Bayesian, 그리고 Bayesian network에서 더 높은 값을 나타내었다. 네트워크 기반 기법을 적용하면 분류에 최적화된 기존 차등 발현 방법의 점수와 많이 달라지게 되는데 실험을 수행한 결과, 동일한 결과를 보여주거나 더 우수한 성능을 나타내었다.

두 번째 실험은 본 연구의 방법으로 해당 질병과 관련된 유전자들을 얼마나 잘 찾을 수 있느냐에 대한 것이다. 미국 국립 생물공학 정보센터(National Center for Biotechnology Information, 이하 NCBI)에 등록된 유전자들 중 전립선 암과 관련된 유전자는 1324개이다. 이 중, 8828개의 유전자로 이루어져 있는 Singh의 마이크로어레이 데이터와 일치하는 유전자는 총 1025개였다. 서열화된 상위 500개의 유전자들 중 전립선 암과 관련된 유전자의 개수를 구하여 Fisher's exact test로 p-value를 계산하였다 (그림 4). 또한 본 연구 방법과 특성 선택 방법인 Chi-square [32], Information gain [33], 그리고 ReliefF [34-36] 들 사이의 전립선 암과 관련된 유전자 개수와 p-value를 비교하였다 (표 3).

서열화된 상위 500개 유전자들 중, 본 연구 방법이 68개로 가장 많은 유전자를 찾아내었다. 또한, p-value도 통계적으로 유의미한 값인 0.02698로 낮은 수치를 기록하였을 뿐만 아니라, Chi-square, Information gain, 그리고 ReliefF 등의 다른 특성 선택 방법과 비교하였을 때도 더 낮은 p-value를 나타내었다.



<그림 4> Fisher's exact test로 구한 본 연구 방법의 p-value 그래프

<표 3> 상위 500개 유전자 중 전립선 암과 관련된 유전자 수와 p-value 비교

	관련된 유전자 개수	p-value
Chi-square	55	0.05419
Information Gain	62	0.04944
ReliefF	64	0.04251
본 연구 방법	**68**	**0.02698**

본 연구 방법에서 찾아낸 유전자 중 나머지 3가지의 방법으로 찾아낸 유전자 중 공통되는 것은 26개였으며, 총 42개의 유전자는 본 연구 방법으로만 찾아낼 수 있었다 (표4). 특히 이 42개의 유전자 중에는 Cancer Genome Project [37] 에서 발표한 전립선 암 종양유전자(Oncogene)인 DDX5가 포함되어 있으며, 여러 가지 암의 종양유전자인 BRCA2, CCND1, JUN, PIM1 등의 유전자 또한 포함되어 있었다. 또한 본 연구 방법으로 찾아낸 유전자 중 하나인 KLK3는 Cancer Genome Project 목록에 포함되어 있지 않았지만, Genome-Wide Association Study(GWAS)를 통해서 전립선 암 종양유전자임이 밝혀진 바 있다 [38]. 이러한 문헌 분석 결과는 본 연구 방법을 이용한 질병 관련 유전자 탐색 방법의 우수성을 간접적으로 보여주고 있다.

<표 4> 본 연구의 방법으로 검색된 유전자 목록

검색된 유전자 목록
SERPINA1, **JUN***, KLK2, GPX3, **KLK3**, FABP5, TPT1, **SEMA3B**, SOD3, PTHLH, HSP90AA1, **ESRRA**, SDC2, **MUC2**, **MAGEA11**, IGFBP2, ATF3, PSMC4, **DDX5**, **CCND1**, RALBP1, ITGAV, **EEF1A1**, SSTR3, **EIF4EBP1**, MAP3K8, **EBAG9**, **PDLIM4**, IL6, **FADD**, **MMP11**, TPM2, LDHA, **EPAS1**, **SEMG1**, **TGFB2**, LOX, ANXA7, **DDC**, **ALDH7A1**, **FLT4**, **ELAC2**, **SKP2**, **ECE1**, **ZFHX3**, **TPD52**, **TYMS**, **SEMG2**, **ID2**, **ENPEP**, **LCN2**, **CST3**, **BRCA2**, CTNNB1, **CXCR4**, **MAGEA4**, **AMH**, **B2M**, **ADRB2**, **PIM1**, **GPX1**, PLAU, PRDX2, ODC1, **WNT5A**, **HLA-A**, **CEBPD**, CSK

* 굵은 글씨는 본 연구에서 제안하는 기법으로만 검색된 유전자임

5. 결론

본 연구 방법은 기존의 차등 발현 유전자 검색 방법에 단백질 상호작용 네트워크 추가적으로 이용함으로써 유전자 간의 상관 관계를 반영하였다. 본 연구에서 제시한 방법은 기존 차등 발현 유전자 검색 방법보다 우수한 높은 분류 정확도와 AUC를 보여주었으며, 암과 관련된 유전자를 탐색함에 있어서도 기존의 방법보다 더 낮은 p-value를 보여줌을 확인할 수 있었다. 차등 발현 분석 기법은 마이크로어레이 데이터 분석 방법 중 가장 기본적이지만 여

전히 가장 많이 사용되고 있는 방법이다. 본 연구에서 제안한 기법은 아직까지도 많이 사용되고 있는 마이크로어레이 데이터뿐만 아니라 앞으로 더욱 많이 사용될 RNA 시퀀싱을 통한 유전자 발현량 분석에도 여전히 효과적으로 적용될 수 있을 것이라고 예상한다.

참고 문헌

[1] Mortazavi,A. et al., “Mapping and quantifying mammalian transcriptomes by RNA-Seq”, Nat. Methods, 5, 621-628, 2008.

[2] Graveley,B.R. et al., “The developmental transcriptome of Drosophila melanogaster”, Nature, 471, 473-479, 2011.

[3] Xu,G. et al., “Transcriptome and targetome analysis in MIR155 expressing cells using RNA-seq”, RNA, 1610-1622, 2010.

[4] H. Chuang, E. Lee, Y. Liu, D. Lee and T. Ideker, "Network-based classification of breast cancer metastasis", Mol. Syst. Biol., 3, 140, 2007.

[5] J. Dutkowski and T. Ideker, "Protein Networks as Logic Functions in Develop-ment and Cancer", PLoS Comp. Bio., 7, e1002180, 2011.

[6] O. Lavi, G. Dror, and R. Shamir, "Network-Induced Classification Kernels for Gene Expression Profile Analysis", Journal of Computational Biology, vol. 19, no. 6, pp. 694-709, 2012.

[7] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria et al., "Dynamic modularity in protein interaction networks predicts breast cancer outcome", Nature Biotechnology, vol. 27, pp. 199-204, 2009.

[8] J. Ahn, Y. Yoon, C. Park, E. Shin and S. Park, "Integrative Gene Network Construction for Predicting a Set of Complementary Prostate Cancer Genes", Bioinformatics, vol. 27, no. 13, pp. 1846-1853, 2011.

[9] G. W. Hatfield, S. Hung and P. Baldi, "Differential analysis of DNA microarray gene expression data", Molecular Microbiology, vol. 47, no. 4, pp. 871-877, 2003.

[10] M. Schena, D, Shalon, R. W. Davis and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray", Science, vol. 270, pp. 467-470, 1995.

[11] Durbin,B.P. et al., “A variance-stabilizing transformation for gene-expression microarray data”, Bioinformatics, 18 (Suppl. 1), S105-S110, 2002.

[12] Ritchie,M.E. et al., “A comparison of background correction methods for two-colour microarrays”, Bioinformatics, 23, 2700-2707, 2007.

[13] Rocke,D.M. and Durbin,B, “Approximate variance-stabilizing transformations for gene-expression microarray data”, Bioinformatics, 19, 966-

972, 2003.

[14] J. Feng, C. A. Meyer, Q. Wang, J. S. Liu, X. S. Liu and Y. Zhang, "GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data", Bioinformatics, vol. 28, no. 21, pp. 2782-2788, 2012.

[15] Cui,X. and Churchill,G.A., "Statistical tests for differential expression in cDNA microarray experiments", Genome Biol., 4, 210, 2003.

[16] Haas,B.J. and Zody,M.C., "Advancing RNA-Seq analysis", Nat. Biotech., 28, 421-423, 2010.

[17] Morozova,O. et al., "Applications of new sequencing technologies for transcriptome analysis", Annu. Rev. Genom. Hum. Genet., 10, 135-151, 2009.

[18] Wall,P.K. et al., "Comparison of next generation sequencing technologies for transcriptome characterization", BMC Genom., 10, 347, 2009.

[19] Wang,Z. et al., "RNA-Seq: a revolutionary tool for transcriptomics", Nat. Rev. Genet., 10, 57-63, 2009.

[20] Bullard,J.H. et al., "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments", BMC Bioinformatics, 11, 94, 2010.

[21] P. Glaus, A. Honkela and M. Rattray, "Identifying differentially expressed transcripts from RNA-seq data with biological variation", Bioinformatics, vol. 28, no. 13, pp. 1721-1728, 2012.

[22] Singh,D. et al, "Gene expression correlates of clinical prostate cancer behavior", Cancer Cell, 1, 203-209, 2002.

[23] Brown,K.R. and Jurisica,I., "Unequal evolutionary conservation of human protein interactions in interologous networks", Genome Biol., 8, R95, 2007.

[24] J. Platt: Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning, 1998.

[25] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, K.R.K. Murthy, "Improvements to Platt's SMO Algorithm for SVM Classifier Design", Neural Computation. 13(3):637-649, 2001.

[26] Trevor Hastie, Robert Tibshirani: Classification by Pairwise Coupling. In: Advances in Neural Information Processing Systems, 1998.

[27] D. Aha, D. Kibler, "Instance-based learning algorithms", Machine Learning. 6:37-66, 1991.

[28] Leo Breiman, "Random Forests", Machine Learning. 45(1):5-32, 2001.

[29] George H. John, Pat Langley, "Estimating Continuous Distributions in Bayesian Classifiers", In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 338-345, 1995.

[30] Heckerman, David, "Bayesian Networks for Data Mining". Data Mining and Knowledge Discovery, 1 (1): 79-119, 1997.

[31] Castillo, Enrique; Gutiérrez, José Manuel; Hadi, Ali S., "Learning Bayesian Networks". Expert Systems and Probabilistic Network Models. Monographs in computer science, pp. 481-528, 1997.

[32] Greenwood, P.E., Nikulin, M.S., "A guide to chi-squared testing", Wiley, New York, 1996.

[33] Koller, D., & Sahami, M., "Toward optimal feature selection", In Proceedings of Thirteenth Conference on Machine Learning, pp. 284-292. Morgan Kaufmann, San Francisco, 1996.

[34] Kenji Kira, Larry A. Rendell: A Practical Approach to Feature Selection. In: Ninth International Workshop on Machine Learning, 249-256, 1992.

[35] Igor Kononenko: Estimating Attributes: Analysis and Extensions of RELIEF. In: European Conference on Machine Learning, 171-182, 1994.

[36] Marko Robnik-Sikonja, Igor Kononenko: An adaptation of Relief for attribute estimation in regression. In: Fourteenth International Conference on Machine Learning, 296-304, 1997.

[37] Cancer Genome Project, http://www.sanger.ac.uk/genetics/CGP

[38] Z. Kote-Jarai et al., "Identification of a novel prostate cancer susceptibility variant in the KLK3 gene transcript", Hum Genet., 129(6):687-94, 2011.

김 현 진

2010년 연세대학교 컴퓨터과학과 졸업(학사)

2010년 - 현재 연세대학교 컴퓨터과학과 통합과정

관심분야 : 바이오인포매틱스, 데이터 마이닝, 텍스트 마이닝, 그래프 마이닝, 데이터베이스

안 재 균

2006년 연세대학교 컴퓨터과학과 졸업(학사)

2009년 연세대학교 대학원 컴퓨터과학과 졸업(석사)

2009년 - 현재 연세대학교 대학원 컴퓨터과학과 박사과정

관심분야 : 바이오인포매틱스, 데이터 마이닝, 데이터베이스 시스템

박 상 현

1989년 서울대학교 컴퓨터공학과 졸업(학사)

1991년 서울대학교 대학원 컴퓨터공학과(공학석사)

2001년 UCLA 대학원 컴퓨터과학과(공학박사)

1991년 - 1996년 대우통신 연구원

2001년 - 2002년 IBM T. J. Watson Research Center Post-Doctoral Fellow

2002년 – 2003년 포항공과대학교
컴퓨터공학과 조교수
2003년 – 2006년 연세대학교
컴퓨터과학과 조교수
2006년 – 2011년 연세대학교
컴퓨터과학과 부교수
2011년 – 현재 연세대학교 컴퓨터과학과 교수
관심분야 : 데이터베이스, 데이터마이닝, 바이오
인포매틱스, 적응적 저장장치 시스템, 플래쉬메
모리 인덱스, SSD