Proceedings of KFIS Autumn Conference 2007 Vol. 17, No. 2.

# 마이크로어레이 데이터의 게놈수준 분석을 위한 퍼지 패턴 매칭에 의한 유전자 필터링 방법

# A gene filtering method based on fuzzy pattern matching for whole genome microarray data analysis

**이선아[1], 이건명[1], 이승주[2], 김원재[3], 김용준[3],배석철[3]**

[1]충북대학교 전기전자컴퓨터공학부
E-mail: kmlee@cbnu.ac.kr
[2]청주대학교 생명유전통계학부
[3]충북대학교 의과대학

**요 약**

생명과학분야에서 마이크로어레이 기술은 세포에서의 RNA 발현 프로파일을 관찰할 수 있도록 함으로써 생명현상의 규명 및 약물개발 등에서 분자수준의 생명현상에 대한 관찰과 분석이 가능해지고 있다. 마이크로어레이 데이터분석에서는 특정한 처리나 과정에서 현저한 특성을 보이는 유전자를 식별하기 위한 분석뿐만 아니라 유전자 전체인 게놈수준에서의 분석도 이루어진다. 최근 유전자의 발현이 다양한 조절, 신호전달 및 대사경로에 의해서 영향을 받고 있다는 관점에서 게놈수준의 분석에 관심이 증가하고 있다. 약물반응 실험에서는 약물에 대한 게놈수준의 발현 프로파일을 관찰하는 것도 많은 정보를 제공할 수 있다. 약물실험에서는 대조군과 실험군들간에 관심있는 상대적인 발현특성을 갖는 유전자군을 전체적으로 추출하는 것이 필요한 경우가 있다. 예를 들면 정상군은 두개의 실험군에 대해서 중간정도의 발현정도를 갖는 유전자군을 식별하는 분석을 하는 경우, 생물학적인 데이터의 특성상 절대값을 비교하는 방법으로는 유용한 유전자들을 효과적으로 식별해 낼 수 없다. 이 논문에서는 정상군과 실험군들의 발현정도값의 경향을 판단하기 위해서 각 유전자에 대해서 집단별 대표값을 선정하여 퍼지집합으로 집단의 값의 범위를 결정하고, 이를 이용하여 특정 패턴을 갖는 유전자들을 식별해내는 방법을 제안하고, 실제 데이터를 통해서 실험한 결과를 보인다.

**Key Words** : 퍼지패턴매칭, 마이크로어레이, 데이터분석, 바이오인포매틱스

## 1. 서 론[1)]

분자생물학의 발전으로 생명현상을 분자수준에서 이해하기 위한 실험, 관찰이 이루어지고 있다. 분자생물학적 실험에서 특정 샘플에서의 유전자 발현 여부를 판정하기 위해서는 DNA가 RNA로 전사될 때의 RNA 양을 측정하는 방법이 일반적으로 사용된다. RNA의 발현량을 측정하는 방법으로 rtPCR 등이 사용되지만 개별 유전자별로 프라이머를 설계하여 실험을 하기 때문에, 많은 수의 유전자에 대해 적용하는 데는 제약이 있다. 마이크로어레이는 유리, 필터 또는 실리콘 판 위에 유전자를 검출할 수 있는 많은 수의 프로우브를 붙여 놓거나 합성하여 놓아서, 동시에 많은 유전자에 대한 발현량을 측정할 수 있도록 한 것이다.[1] 마이크로어레이 기술의 발전에 따라 현재 동시에 4만여 유전자의 발현을 동시에 측정할 수 있는 제품[6]이 출현하는 등, 대규모 유전자의 발현정도 측정이 가능해지고 있다. 마이크로어레이는 동시에 많은 양의 데이터를 생성하기 때문에, 이에 대한 효과적인 분석기술이 필요하다. 마이크로어레이는 약물효능 분석 분야에서도 사용되고 있다. 동물 실험 등에서 약물을 투약한 것과 하지 않는 샘플들에 대해서 마이크로어레이 분석하여, 약물에 반응한다고 판단되는 단백질을 추정하고, 약리메카니즘을 규명하기 위한 연구를 하고 있다.

생체 내의 대사경로, 신호전달경로, 유전자 발현조절경로는 네트워크 구조를 갖기 때문에, 약물의 투약에 따라 큰 차이를 보이는 유전자가 발견되기도 하지만, 여러 유전자 군이 특정 패턴을 보이는 형태로 변할 수 있다. 따라서,

1) 이 논문는 2007년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (지방연구중심대학육성사업/충북BIT연구중심대학육성사업단)

약물에 대한 영향을 분석할 때 마이크로어레이를 사용하는 경우에는 전체 유전자 집합, 즉 게놈 전체에 대한 관점에서 분석하는 것이 바람직한 경우도 있다. 예를 들면, 항암 후보약물의 효과를 동물실험을 통해서 확인하는 실험인 경우, 발암물질을 사용하여 실험동물에 암을 유발시키면서 항암물질의 효능을 관찰하게 된다. 항암물질은 정상 동물의 세포에 대해서는 영향을 주지 않으면서, 발암물질 투여에 따라 암이 발생할 수 있는 동물의 세포에 대해서는 암이 발생이 억제되는 효과를 보이는 것이어야 한다. 이러한 항암물질의 효능 분석을 위한 마이크로어레이 실험인 경우에는 게놈 전체에 대해서 발암물질, 항암후보 물질의 투약 여부에 따른 실험동물 세포의 암의 유발 여부와 함께 해당 세포의 마이크로어레이를 통한 유전자 발현정도를 분석함으로써, 항암 후보물질의 안전성, 항암 후보물질에 영향을 받는 유전자 중에서 암유전자(oncogene), 암억제유전자(suppressor gene) 등을 추정하고, 추가적인 분석을 할 수 있다. 이와 같은 분석을 위해서는 유전자의 발현정도를 특정 기준에 따라 나누어진 집단별로 비교하는 것이 필요하다. 생물체의 특성상 발현정도는 확률적인 분포를 가지고 나타난다고 봐야 하기 때문에, 절대적인 비교 연산을 하는 것이 곤란하다. 이러한 비교를 위해 퍼지비교를 하여 특정 퍼지 패턴을 만족하는 유전자를 추출하는 방법을 제안한다.

2절에서는 마이크로어레이를 분석하는 전형적인 방법에 대해서 소개하고, 3절에서는 퍼지 소속함수를 사용하여 집단간의 발현 패턴을 지정하여 이를 만족하는 유전자를 추출하는 방법을 제안하고, 4절에서는 이를 적용한 실험결과의 예를 보이고, 5절에서 결론을 맺는다.

## 2. 마이크로어레이 데이터분석

마이크로어레이 데이터는 여러 샘플에 대한 실험결과를 비교하는 것이 일반적이기 때문에 2차원 행렬형태로 주어지는데, 각 행은 하나의 유전자에 대응하고, 각 열은 하나의 샘플에 해당하고, 원소는 해당 유전자의 샘플에서의 발현정도를 나타낸다. 마이크로어레이는 하나의 형광염색을 이용하는 1채널 방식과, 두가지 형광염색을 이요하는 2채널 방식이 있다. 유전자의 RNA 발현량을 측정하기 위해, 샘플 세포로부터 RNA를 분리하고, 이에 대한 cDNA를 만들고 이를 PCR(polymer chain reaction)을 이용하여 증폭을 하고 형광염색하여, 이를 마이크로어레이 칩에 올려주면, 베이스 상보결합에 의해 마이크로어레이 칩에 부착되게 된다. 칩에 부착된 cDNA를 형광 스캔을 통해 읽어들이면, 이미지 형태의 데이터를 얻게 된다. 이러한 데이터에 대한 영상처리를 통해 부착정도에 대응하는 수치정보를 읽어내어 이를 마이크로어레이 데이터로 제공하게 된다. 이러한 일련의 실험은 실제 시험관을 이용한 실험이면서, 스캔 및 영상처리를 해야 하기 때문에 의도하지않은 변이가 포함될 수 밖에 없다. 따라서 이러한 변이를 최소화하기 위해서, 데이터 분석 이전 단계에서 전처리로서 정규화(normalization)를 시켜야 한다. 정규화과정에서는 스캔된 영상으로부터 수치값을 변환과정에서 데이터의 신뢰도 때문에 값이 주어지지 않은 경우도 있고, 발현정도의 값의 범위가 유전자별로 제각각 이기 때문에 이를 처리해줘야 한다.

정규화 등의 전처리가 완료되면, 2차원 행렬로 주어진 마이크로어레이 데이터에 대한 분석을 하게 된다. 분석의 관점에 따라 여러 가지 분석방법이 사용되게 된다. 대표적인 기본 분석으로 군집분석(cluster analysis)이 많이 사용되고, 이외에도 분류분석, 차원축소를 통한 정보 추출을 위한 PCA 등 여러 분석이 사용된다. 군집분석에서는 유전자들에 대한 클러스터 트리 구성뿐 만아니라 경우에 따라서는 샘플들에 대해서도 클러스터 트리를 구성하게 할 수 있다. 마이크로어레이 분석 결과로부터 분석자들은 관심있는 추가적인 분석을 수행하게 된다. 군집분석에서는 인접한 행간에, 또는 인접한 열간의 유사성이 크도록 열 및 행을 재배치하게 되기 때문에, 유전자의 개수가 많은 경우 관심있는 패턴을 갖는 유전자집단을 효과적으로 찾기 위해서는 추가적인 분석 방법이나 도구가 필요하다.

약물후보물질 실험인 경우에는 약물에 대한 효과가 있는 실험집단과 그렇지 않은 집단, 약물에 대한 처리를 하지 않은 집단의 샘플이 발생한다. 또한 항암 치료 후보물질의 실험인 경우에는 발암물질에 노출된 집단에 대해서 후보물질을 투여해서 효과가 있는 집단과 그렇지 않은 집단을 비교분석을 한다. 항암 후보물질의 경우에는 약물 투여후에도 정상 샘플과 비교하여 분자생물학적 수준, 즉 마이크로어레이 데이터 수준에서는 통계적으로 유의한 차이가 없어야 한다. 반면, 발암물질에 노출된 실험 집단에 대해서는 효과가 있는 것이 분명히 나타나야 바람직하지만, 어떤 경우에는 효과가 있고, 그렇지 못할 수도 있다. 따라서, 게놈 수준에서 효과가 있는 경우 역할을 하는 발암억제 유전자들에 대한 분석과 효과가 없는 경우 암 유전자들에 대한 분석을 위해 이러한 유전자를 필터링해서 찾아내는 것이 필요하다. 암억제

유전자 집단이 될 수 있는 것은 발암물질과 함께 투여되어, 암의 억제 효과가 있는 샘플들에서는 효과가 없는 샘플들에서 보다 높이 발현되면서, 정상 샘플에서 보다도 높은 발현정도를 보이는 것들이다. 한편, 암유전자가 될 수 있는 것들은 발암물질과 후보약물이 함께 투여됐을 때, 효과가 없는 집단의 발현정도가 효과가 있는 집단의 발현정도보다 높으면서, 정상 집단의 발현정도보다는 높은 것들이다.

## 3. 퍼지 페턴매칭에 의한 유전자 필터링

### 3.1 암유전자 후보집단 및 암억제 유전자 후보집단

분석 대상는 항암 후보물질의 효능을 보는 마이크로어레이 분석 실험으로, 정상집단에 대한 샘플(N)과, 발암물질과 항암 후보물질을 동시에 투약한 실험한 동물 샘플이 확보되고, 이들 샘플에 대한 마이크로어레이 실험을 통해 유전자 발현 정보가 얻어진 경우를 전제로 한다. 발암물질(O)과 항암 후보물질(S)의 동시투약 집단에 대한 결과로 효과가 있는 집단((O+S)P)과 효과가 없는 집단((O+P)N)으로 마이크로어레이 데이터를 구별한다. 이 때, 게놈 전체에 대해서 암억제 유전자 및 암유전자 후보가 될 수 있는 것은 유전자의 발현정도값이 아래 같은 특성을 가지는 것이다.

■암억제 유전자 후보의 집단간의 발현특성

$$N < (O+P)N < (O+P)P \quad (1)$$

■암유전자 후보의 집단간의 발현특성

$$N < (O+P)P < (O+P)N \quad (2)$$

생물체에서는 유전자의 발현이 다양하고 복잡한 조절, 대사, 신호전달 네트워크에 의해 영향을 받기 때문에, 위와 같은 기준에 의한 후보 유전자의 선정 방법과 같이 이상적인 형태로 나타나지 않고 확률적인 특성을 가지고 있다고 전제하고 분석하는 것이 필요하다. 즉, 발현정도값의 절대적인 비교만으로 곤란하고, 일부에 대해서는 이러한 대소관계를 만족하지 못하더라도 후보로 선정할 수 있는 방법이 필요하다. 이러한 요구에 부응하기 위한 방법으로 소속함수를 이용하여 정의한 퍼지패턴을 이용하여 후보 유전자를 선정하는 방법을 제안한다.

### 3.2 유전자 선정을 위한 퍼지패턴 표현

마이크로어레이 데이터로부터 식 (1), (2)와 같은 조건을 만족하는 유전자를 선택하기 위해 마이크로어레이 데이터값의 범위 내에 각 집단에 대해서 해당 집단의 유전자 발현정도값을 아우르는 퍼지집합에 대한 소속함수를 정의한다. 이러한 소속함수를 정의하기 위해서 다음과 같은 방법을 이용한다.

단계 1. 각 유전자 $g_i$별로 샘플 집단 $S_j$ 각각에 대해서 유전자 발현정도값 $e_{ik}$의 평균 $m_{ij}$을 계산한다.

$$m_{ij} = \frac{\sum_{e_{ik} \in S_j} e_{ik}}{|S_j|} \quad (3)$$

단계 2. 각 유전자 $g_i$별로 모든 발현정도값을 오름차순으로 정렬한다. $|S|$는 전체 샘플갯수를 나타낸다.

$$EL_i = (e_{i(1)}, e_{i(2)}, \cdots, e_{i(|S|)}) \quad (4)$$

단계 3. 각 샘플집단 $S_j$별로 평균값 $m_{ij}$을 기준으로 정렬된 값을 왼쪽으로 따라가면서, 다른 집단의 값이 나오지 않으면서 $S_j$에 속하는 가장 작은 값 $lv_{ij}$을 선택한다.

단계 4. 각 샘플집단 $S_j$별로 평균값 $m_{ij}$을 기준으로 정렬된 값을 오른쪽으로 따라가면서, 다른 집단의 값이 나오지 않으면서 $S_j$에 속하는 가장 큰 값 $rv_{ij}$을 선택한다.

단계 5. 각 유전자별 $g_i$별로 샘플 집단 $S_j$ 각각에 대해서 다음과 같이 퍼지집합 $TR_{ij} = Trap(lb_{ij}, lv_{ij}, rv_{ij}, rb_{ij})$을 정의한다. 편의상 식 (1)과 (2)와 같은 조건을 만족하는 유전자를 찾을 때, 마이크로어레이 데이터에서 집단의 배치 순서가 기대하는 크기순서로 되어 있다고 가정한다. 샘플 집단 $S_1$, 즉 가장 작은 값을 갖는 것이 기대되는 집단과 가장 큰 값을 갖는 것이 기대되는 집단인 $S_L$인 경우를 제외하면 $TR_{ij} = Trap(lb_{ij}, lv_{ij}, rv_{ij}, rb_{ij})$은 사다리꼴 퍼지숫자를 나타낸다. $S_1$인 경우에는 $lb_{i0} = -\infty$인 형태로, 즉 소속함수의 왼쪽 부분의 값이 모두 1인 형태의 소속함수가 되고, $S_L$은 $rb_{iL} = \infty$인 형태로, 소속함수의 오른쪽 부분의 값이 모두 1인 형태를 나타낸다. $lv_{ij}, rv_{ij}$는 단계 3, 4에서 구한 값이다. $lv_{ij}$의 값은 구간 $[m_{i(j-1)}, lv_{ij}]$의 값을 선택하고, $rv_{ij}$의 값은 구간 $[rv_{ij}, m_{i(j+1)}]$의 값을 선택한다. 그림 1은 3개의 집단에 대한 식 (1)이나 (2)와 같은 특성을 퍼지 패턴으로 표현한 예이다.

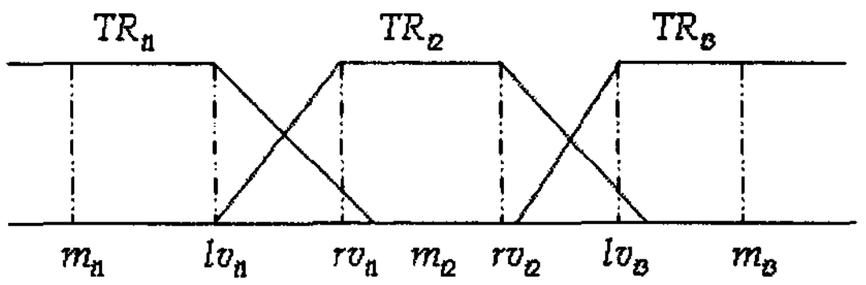


그림 1. 퍼지 패턴을 표현하는 소속함수

### 3.3 퍼지 패턴을 이용한 유전자 선정

게놈 전체에 대한 마이크로어레이 데이터로부터 퍼지 패턴으로 표현된 성질을 만족하는 유전자를 선택할 때는 다음과 같은 과정을 따른다.

단계 1. 각 유전자 $g_i$에 대해서 샘플 집단 $S_j$별로 퍼지집합의 소속함수 $TR_{ij}$에 대한 유전자 발현정도값 $e_{ik}$의 소속정도 $md_{ik}$를 계산한다.

$$md_{ik} = \mu_{TR_{ij}}(e_{ik}),\ \text{where}\ e_{ik} \in S_j \qquad (5)$$

단계 2. 유전자 $g_i$에 대해서 샘플 집단 $S_j$별로 소속함수의 평균 $av_{ij}$을 계산한다.

$$av_{ij} = \frac{\sum_{e_{ik} \in S_j} md_{ik}}{|S_j|} \qquad (6)$$

단계 3. 유전자 $g_i$의 각 샘플 집단 $S_j$별 평균 소속함수 값 $av_{ij}$이 정해진 임계값 $\theta$이상이면, 이를 주어진 패턴을 만족하는 유전자로 선택한다.

위와 같은 방법을 주어진 마이크로어레이 데이터에 있는 모든 유전자에 대해서 적용하면, 주어진 패턴을 만족하는 유전자를 모두 선정할 수 있다.

## 4. 실험 및 고찰

제안한 방법에 대한 유용성을 보이기 위해서 실제 효능실험 중인 항암물질의 마이크로어레이 데이터 분석에 퍼지패턴을 이용한 유전자 선정 방법을 적용하였다. 그림 2는 주어진 원래 데이터를 전처리한 후에 군집분석할 결과의 일부를 보인 것이다.

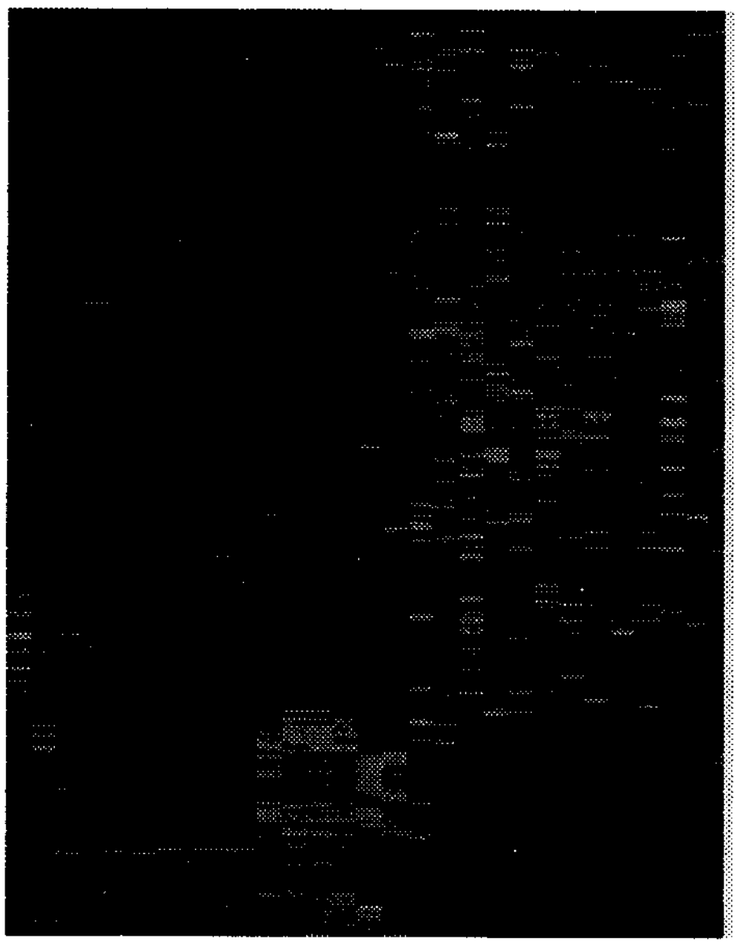

그림 2. 마이크로어레이 데이터의 군집분석 예

그림 3은 그림 2의 데이터에 대해서 제안한 방법에 따라 퍼지 패턴을 정의하여 주어진 성질을 만족하는 유전자를 게놈 전체로부터 추출한 결과를 보인 것이다. 그림에서 보는 바와 같이 집단별로 발현정도가 유사한, 그렇지만 다른 집단과의 약간의 중첩이 있는 것을 허용하는 유전자를 효과적으로 추출했음을 확인할 수 있다.

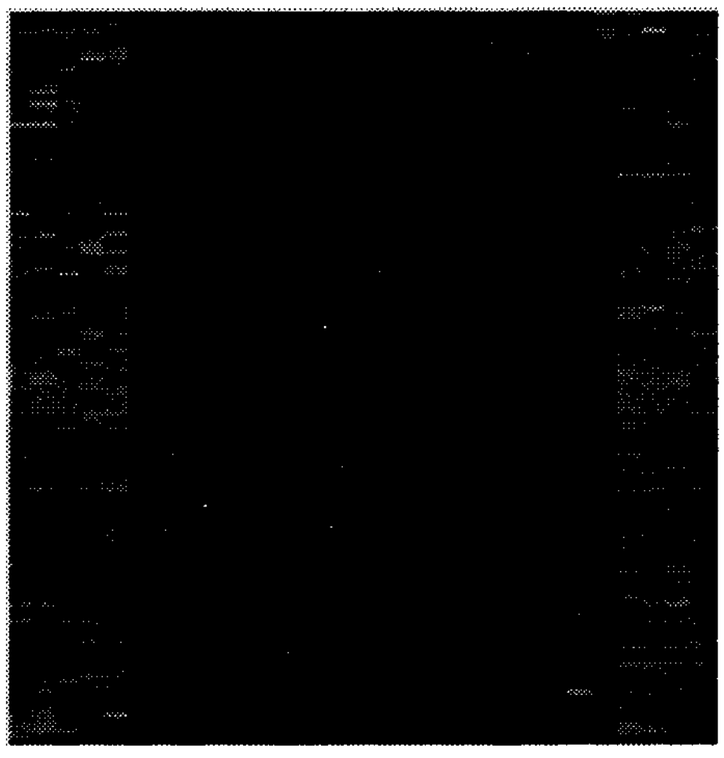

그림 3. 퍼지 패턴을 이용한 유전자 선택결과

## 5. 결론

이 논문에서는 마이크로어레이 데이터분석에서 게놈 전체에 대해서 약물효능 분석과 같은 경우에 필요한 특정 패턴을 보이는 유전자를 효과적으로 선택하기 위한 방법으로 퍼지 패턴을 이용하는 방법을 제안하였다. 제안한 방법은 실제 항암 후보물질의 실험결과에 대한 분석에 효과적으로 적용될 수 있을 확인하였다. 향후 발현정도의 대소비교뿐만 아니라 상관관계를 반영한 패턴을 만족하는 유전자 선택방법에 대한 연구를 수행할 예정이다.

## 참 고 문 헌

[1] S. Draghici,"Data Analysis Tools for DNA Microarrays", Chapman & Hall/CRC, 2003.

[2] D. W. Mount, “Bioinformatics: Sequence and Genome Analysis,” Cold Spring Habor Lab Press, 2004.

[3] G. B. Forgel, D. W. Corne, “Evolutionary Computation in Bioinformatics,” Morgan Kaufmann Publishers, 2003

[5] W. L. Martinez, A. R. Martinez, “Exploratory Data Analysis with MATLAB,” Chapman&Hall/CRC, 2005.

[6] Illumina, “BeadStudio Gene Expression Module User Guide,” Illumina, 2006