한국지능시스템학회 논문지 2008, Vol. 18, No. 1, pp. 14-19

기호코딩을 통한 마이크로어레이 데이터의 추이 패턴 추출

Trend Pattern Extraction from Microarray Data with Symbolic Encoding

이선아[1], 이건명[1], 김원재[2]
Sun A Lee, Keon Myung Lee, Wun Jae Kim

[1]충북대학교 전기전자컴퓨터공학부
[2]충북대학교 의과대학

요 약

대규모로 유전자 발현정도를 동시에 측정하는 마이크로어레이 실험은 많은 양의 데이터를 생성하기 때문에, 자동화된 효과적인 분석기법이 필요하다. 이 논문에서는 약물의 영향 분석을 위해 약물의 투여량 및 투여후의 시간대별로 샘플을 추출하여, 마이크로어레이를 이용하여 유전자의 발현량을 분석하는 경우에, 약물에 대해서 반응하는 유전자를 추출하는 데이터 마이닝 기법을 제안한다. 제안한 방법에서는 유전자의 발현정도값을 이전 비교대상의 값을 기준값으로 하여 증가, 감소, 답보에 해당하는 기호로 매핑하여, 분석자가 원하는 패턴을 보이는 유전자를 추천한다. 한편, 유전자의 상호간에 많은 영향을 주고 받기 때문에 특정 약물을 투여할 때, 이에 직접적인 영향을 받는 것도 있지만, 이와는 전혀 상관없이 동작하는 것도 있기 때문에, 제안한 방법에서는 이러한 약물 투여와 유의성이 있을 가능성이 있는 유전자만을 전처리과정을 통해서 필터링하는 기법을 활용한다.

Key Words : 바이오인포매틱스, 마이크로어레이, 데이터분석, 추이 분석, 데이터마이닝

1. 서 론

생명체의 모든 유전정보는 염색체 내의 DNA 서열에 저장되어 있고, 이들 서열에서 생체의 조절 메카니즘에 따라 특정 유전자들이 RNA로 전사되어 발현되어 아미노산 서열을 조합하여 단백질을 만들어낸다.[1] 단백질은 생명체의 구성, 대사, 면역, 발병, 번식 등 대부분의 주요 생명현상에 관여하는 핵심 고분자 물질이다. 유전자 지도를 재구성하려는 지놈 프로젝트 이후로 많은 생명과학자들은 유전자의 기능을 예측하는데 많은 노력을 집중하고 있다.

유전자의 기능 예측 및 유전자의 발현과 질병의 발병, 소멸, 전이 등의 관계를 분석하는 것은 질병의 치료, 예후 추정, 신약 개발 등의 분야에서 그 중요성이 커지고 있다.[2] 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. 마이크로어레이(microarray)는 유리, 필터 또는 실리콘 판 위에 유전자를 검출할 수 있는 많은 수의 프로우브(probe)를 붙여 놓거나 합성하여 놓아서, 동시에 많은 유전자에 대한 발현량을 측정할 수 있도록 한 것이다. 마이크로어레이 기술의 발전에 따라 현재 동시에 4만여 유전자의 발현을 동시에 측정할 수 있는 제품[5]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 약물의 투약량을 달리한 여러 군에 대해서 투약시간이 다른 샘플들에 대해서 마이크로어레이 분석하여, 약물에 반응한다고 판단되는 단백질을 추정하고, 약리메카니즘을 규명하기 위한 연구를 하고 있다. 현재 사용되는 마이크로어레이는 하나의 샘플에 대해서 수만여종 이상의 유전자에 대한 발현을 측정하는 경우도 있기 때문에, 유의한 유전자를 추정하는 효과적인 방법이 필요하다. 이 논문에서는 약물효능 분야의 실험에 마이크로어레이를 사용하는 경우의 약리반응에 유의한 반응을 보이는 유전자를 찾는데 적용할 수 있는 기호코딩을 통한 데이터마이닝 기법을 소개한다.

접수일자 : 2007년 11월 20일
완료일자 : 2007년 12월 25일
이 논문은 2007년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (지방연구중심대학육성사업/충북BIT연구중심대학육성사업단)

2. 관련 연구

2.1 약물반응 분석에서의 마이크로어레이

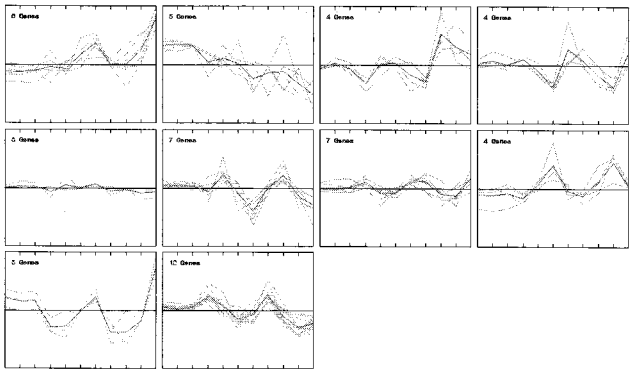
새로운 약물은 개발하는 과정에서는 많은 기반연구가 수행된다. 신약개발은 일반적으로 타켓 식별(target identification)을 시작으로, 타켓에 대한 검증(target validation), 선도물질 도출 및 최적화(lead identification and optimization)을 거쳐 후보약물이 개발된다. 이러한 후보약물은

전임상 개발(preclinical development)을 거쳐, 1상, 2상, 3상 임상실험(phase I, II, III clinical trials)을 통해서, 효능이 입증되면 신약으로 출시하게 된다. 선도물질 도출 및 최적화 과정에서는 효능에 관련된 약역학(pharmacodynamics)에 대한 연구, 안전성, 흡수(absorption), 확산(distribution), 대사(metabolism), 배설(excretion) 등에 대한 역동력학(pharmocokinetics)에 대한 연구가 수행된다. 후보약물의 효능 평가를 위해서 질환동물모델을 통한 실험을 하여, 효능 및 안전성 등에 대한 데이터를 수집하여 분석하게 된다. 동물실험에 대한 분자생물학적 분석을 위해 마이크로어레이를 활용하기도 한다. 후보약물의 생체내에서의 효과를 분석하기 위해, 분자생물학적인 관점에서 어떠한 유전자가 약물에 반응하는지 확인하고, 이들에 의한 대사를 추정하는 것은 약물 개발에서 매우 의미있는 작업이다. 약물에 반응하는 유전자의 발현패턴은 약동력학적인 관점에서도 유용한 정보를 제공한다. 마이크로어레이는 동시에 많은 수의 유전자 발현을 측정할 수 있는 도구인데, Affimtrix, Illumina 등의 회사에서 동시에 수만종의 유전자의 발현 정도를 측정할 수 있는 마이크로어레이를 제공하고 있다. 약물에 대한 반응을 확인하기 위해서, in vivo나 in vitro 실험에서는 약물의 투약량을 바꾸어가면서 실험을 하고, 시간의 경과에 따른 효과를 확인하기 위해 시간대 별 샘플을 확보하여 실험을 한다. 이와 같이 다수의 실험에서 각각 수만종의 유전자에 대한 발현 데이터가 발생하는 상황에서 유의한 유전자를 수작업으로 추출하는 것은 비용이 많이 들 수 있다. 이 논문에서는 이러한 상황에 적용할 수 있는 데이터분석 기법으로서, 유전자의 발현정도를 기호로 코딩하여, 특정 패턴을 보이는 약물을 효과적으로 추출하는 방법을 제안한다.

2.2 추이 분석 방법

약물 실험에서는 약물의 투여량과 경과시간에 따른 효과를 측정하는 것이 필수적인 요소의 하나이다. 마이크로어레이 데이터로부터 유전자에 대한 RNA 발현 농도를 측정하는 경우, 약물의 영향을 분자생물학적 수준에서 관찰할 수 있다. 추이분석을 위해 사용될 수 있는 대표적인 것으로 비지도(unsupervised) 학습 방법인 간주되는 SOM(self-organizing map), k-means 알고리즘 등[5]이 있다. SOM은 신경회로망 모델의 하나로서, 데이터를 크기가 1이 되도록 정규화한 다음, 유사한 패턴을 보이는 샘플들을 군집화시키는 역할을 한다.k-means 알고리즘은 군집의 집단을 중심으로 최근접 집단으로 데이터를 분류하고, 분류된 데이터들로부터 새로운 중심을 결정하는 방법을 반복하여 군집화를 한다. k-means 알고리즘을 시간대별 마이크로어레이 데이터에 대해서 적용하면, (그림 1)과 같은 형태로 유사한 패턴을 보이는 샘플끼리 군집화하는 것이 가능하다.

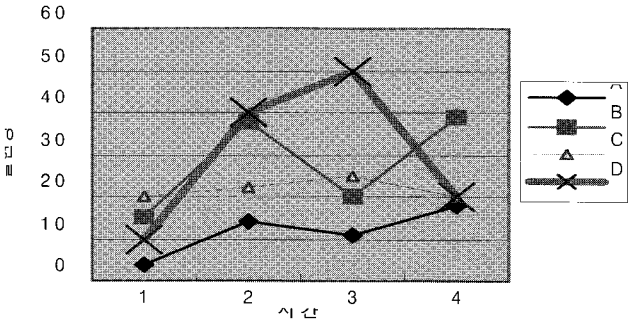
마이크로어레이 데이터 분석에서는 시간의 경과 또는 약물투여량의 증가에 따른 RNA 발현량의 특정 증가감소 추이 패턴을 보이는 것을 식별해야 하는 경우가 있다. (그림 2)은 4개의 샘플에 대한 RNA 발현량을 시간대별로 측정한 것을 예시한 것으로, 샘플 A와 B는 증가하다가 감소하고, 다시 증가하는 패턴을 보이고, C와 D 샘플은 시간의 경과에 따라 어느 정도 증가하다가 감소하는 패턴을 보인다.



(a)

그림 1. k-means 알고리즘에 의한 추이 데이터의 군집화

	t_1	t_2	t_3	t_4
A:	4	14	11	18
B:	15	38	20	39
C:	20	22	25	20
D:	10	40	50	20



(b)

그림 2. 시간경과에 따른 발현 프로파일 예

이러한 추이를 갖는 유전자를 찾아내는데 기존의 k-means 알고리즘, SOM 등에서 사용하는 거리 기반 방법으로는 기대하는 결과를 얻기 어려운 경우가 있다. (표 1)은 k-means 등에서 사용하는 유클리드(Euclid) 거리를 이용하여 데이터간 거리를 나타낸 것이고, (표 2)는 SOM 등에서 사용하는 데이터를 크기가 1이 되도록 정규화한 후에 이들 간의 거리를 나타낸 것이다. 증감추이 패턴을 분석할 때는 샘플 A와 B가 동일한 패턴으로 분류되고, 샘플 C와 D는 또 다른 패턴으로 분류되는 것을 기대한다. 그런데 (표 1)의 거리 정보를 사용하면, 샘플 A, B, C가 동일한 패턴으로 분류될 수 있고, (표 2)를 사용하면 샘플 C가 D보다 B에 더 가깝게 판단됨으로, C와 D가 다른 패턴으로 분류될 수 있다. 이와 같이 단순히 거리 정보를 이용하여 패턴을 분류하는 것은 마이크로어레이 데이터의 추이분석에서 요구되는 패턴을 기대하는 대로 식별할 수 없을 수도 있다. 따라서 이 논문에서는 거리 정보를 이용하는 대신에, 이전 시점의 발현량의 상대적인 증감여부 정보를 이용하는 추이 패턴을 식별하는 방법을 제안한다.

표 1. (그림 2)의 데이터간 절대거리

	A	*B*	*C*	*D*
A	0	1219	520	2237
B	1219	0	667	1290
C	520	667	0	1049
D	2237	1290	1049	0

표 2. (그림 2)의 데이터를 정규화한 후 데이터간 거리

	A	*B*	*C*	*D*
A	0	0.028	0.173	0.263
B	0.028	0	0.154	0.302
C	0.173	0.154	0	0.157
D	0.263	0.302	0.157	0

3. 기호코딩 기반의 추이 패턴 추출

어떤 값의 증가감소 추이에 대한 패턴을 식별하고 추출할 때, 거리기반의 k-means 알고리즘, SOM 등을 사용하는 경우 (그림 2)의 예에서 확인한 것처럼 추이가 다른 데이터들이 동일한 패턴으로 분류될 수 있다. 약물투약 효과 분석 등과 실험에서 추이를 확인하기 위해서는 시간 경과 또는 투약량에 따른 영향의 증가 및 감소 경향에 관심을 갖게 되는 경우가 많으므로, 증가감소 추이 패턴을 추출하는 방법으로 데이터 획득과정에서 인접하는 값 간의 상대적인 증가감소를 기호로 코딩하는 방법을 제안한다. 기준을 정하여 현재 값이 이전값보다 증가하면 H로 현재값을 코딩하고, 이전값보다 작으면 L로 코딩하고, 큰 차이가 없으면 E로 코딩하는 것이다. (그림 2)의 데이터를 이와 같은 방법으로 코딩하면 다음과 같은 코드화된 데이터를 얻게 된다.

A : H L H
B : H L H
C : H H L
D : H H L

위에서 보는 것과 같이 샘플 *A*와 *B*, *C*와 *D*는 각각 HLH 및 HHL 패턴을 갖는 동일 한 것으로 분류되어 식별된다.

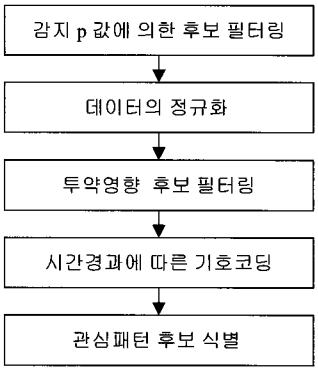


그림 3. 약물투약 마이크로어레이 추이패턴 분석과정

마이크로어레이 데이터의 고유한 특성 때문에 위와 같은 코딩 방법을 적용하기 위해서는 전처리과정 및 코딩에 대한 적절한 고려가 필요하다. (그림 3)은 이 논문에서 제안하는 약물투약 마이크로어레이 데이터에 대한 추이패턴 추출을 위한 과정을 도시한 것이다.

3.1 감지 p 값에 따른 후보 필터링

마이크로어레이 데이터는 (그림 4)와 같이 많은 수의 유전자에 대한 발현량에 대한 수치값 및 관련정보로 구성된다. 약물에 유의한 반응을 보이는 후보를 추출하기 위해서는 가능하면 전처리과정에서 연관성이 없을 것으로 판단되는 많은 유전자를 고려대상에서 제외하는 것이 필요하다. 이를 위해서는 약물 효능실험에 대한 특징과 분석배경지식을 활용하는 것이 필요하다.

	A			
TargetID	AVG_Signal	BEAD_STDERR	Avg_NBEADS	Detection Pval
18S_rRNA_X00686_...	39824.0	1341.506	42	0.00000
18S_rRNA_X00686_...	14562.9	350.501	37	0.00000
18S_rRNA_X00686_...	21214.8	411.096	64	0.00000
23334588_104_rc-5	335.5	10.456	47	0.00745
9626100_15-5	705.7	52.353	45	0.00000
9626100_224-5	2855.2	224.211	49	0.00000
9626123_31_rc-5	280.1	9.839	35	0.01925
9626953_200-5	1010.5	63.236	43	0.00000
9626962_229-5	416.6	28.213	37	0.00248
9629514_325-5	481.7	15.824	56	0.00124
9845300_5100-5	735.7	31.945	54	0.00000
GI_10946639-5	1507.9	35.371	54	0.00000
GI_13507625-5	19153.9	435.603	47	0.00000
GI_14149646-5	52437.2	1546.035	49	0.00000
GI_18079338-5	1700.4	42.279	50	0.00000
GI_18093091-5	650.2	16.297	48	0.00000
GI_19111151-I	1497.4	51.097	38	0.00000
GI_19527057-5	886.9	28.643	52	0.00000
GI_19527195-5	10404.2	308.671	44	0.00000
GI_20347917-5	299.1	11.766	45	0.01491
GI_20799906-5	2802.5	80.273	48	0.00000
GI_20819699-5	1055.0	41.601	38	0.00000
GI_20825167-5	7129.0	178.333	49	0.00000

그림 4. 마이크로어레이 데이터의 예

약물반응 분석에 유의한 후보 유전자 선정을 위해서 마이크로어레이 실험에서 측정치에 대한 신뢰도가 떨어지는 것을 우선적으로 제외한다. 마이크로어레이 스캔 도구는 마이크로어레이에서 각 유전자에 대응하는 위치로 부터 발현정도에 대한 값을 이미지 형태로 읽어들인다. 이때 (그림 4)에서 보는 바와 같이 측정치와 함께 측정치에 대한 정확도를 감지 p 값(Detection p value)이라는 형태로 제공한다. 이 값이 클수록 정확도가 의심되기 때문에, 이 값이 특정값(예, 0.05) 이상이면 추후 분석에서 고려할 유전자 집단에서 배제한다.

3.2 데이터의 정규화

마이크로어레이 분석에서는 특정 유전자의 발현량이 비정상으로 큰 값을 갖는 경우도 있고, 발현량의 값의 범위가 연전자 간에 차이가 많이 날 수 있기 때문에 정규화(normalization)를 하는 것이 필요하다. 가장 쉽게 적용가능한 정규화 방법은 발현량의 값을 정렬한 후, 특정 순위 이상의 값에 대해서 일정한 값으로 고정하는 것이다. 한편, 로그(log) 함수를 적용하여 발현량의 값이 매우 큰 부분이나 지나치게 작은 부분에 대한 민감도를 떨어뜨리는 방법을 사용하기도 한다.

3.3 투약영향 후보 필터링

분자생물학차원에서 유전자를 살펴보면 생명체의 유지를

위한 다양한 상호작용에 의해서 유전자의 발현이 상호영향을 주기 때문에 유전자의 발현 양태는 매우 다양하다. 이러한 복잡한 동적인 상황에서 약물에 영향을 받은 유전자의 후보를 선택하는 것은 쉬운 일이 아니다.

제안한 방법에서는 정상(control) 집단에 대한 샘플에서 시간이 경과하여도 평균적으로 일정한 발현정도를 유지하는 유전자를 약물에 영향을 받을 수 있는 집단의 후보로 선택하여 추가적인 분석을 한다. 이러한 유전자를 선택하기 위해서, 다음과 같은 과정을 통해서 추후 분석할 유보 유전자들을 선정한다.

단계 1. 정상 집단 샘플에 대해서 각 유전자별로 평균 발현량 CTR_{avr}을 계산한다.
단계 2. 정상 집단에 대해서 각 유전자 별로 최대 발현량과 최소 발현량의 차이 CRT_{int}를 계산한다.
단계 3. 변화량이 적은 유전자들을 선택하기 위해서, 유전자별로 발현량의 변화율을 $CTR_{var} = CTR_{int}/CTR_{avg}$를 이용하여 계산한다.
단계 4. CTR_{var}기준으로 유전자를 오름차순의 정렬한다.
단계 5. 정렬된 유전자 리스트에서 CTR_{var}값이 일정값 (예, 0.05)이하인 것들을 추후 고려할 유전자 집단으로 선정한다.

이 필터링 과정을 거치게 되면, 고려할 유전자가 수가 수만 종에서 수백 종까지 줄어 들 수 있게 된다.

3.4 시간경과 또는 투약량에 따른 증감패턴의 기호코딩

수치로 표현된 마이크로어레이 데이터를 추이를 표현하는 기호로 코딩하기 위해 다음과 같은 방법을 사용한다. 우선 각 유전자별로 샘플전체에 대한 평균 발현정도 ALL_{avg}를 결정한다. 이 평균 발현정도값을 기준으로 직전 비교대상 값 ($preceed_{val}$) 대비 현재 값(cur_{val})을 보고 현재값의 상태를 나타낸다. 즉, 직전 비교값보다 현재값이 일정수준이상 크면, 현재값을 'H'로 코딩하고, 비슷하면 'E'로, 일정수준보다 작으면 'L'로 변경하고 코딩한다. 시간 경과에 대한 추이 패턴 추출인 경우에는 이전 비교값이 현재 시점 바로 이전의 측정값이 되고, 투약량에 따른 비교인 경우에는 비교 투약량보다 적게 투약된 직전 투약량에 대한 측정값이 된다.

다음은 위의 H, E, L 기호를 사용하는 코딩 방법을 예시한 것이다.

if (abs($< eed_{val}$-cur_{val}) < ALL_{avg}*α), 'E'로 코딩
if ($< eed_{val} > cur_{val}$+ALL_{avg}*α), 'L'로 코딩
if ($< eed_{val} < cur_{val}$+ALL_{avg}*α), 'H'로 코딩

그림 5. 기호코딩 규칙

(그림 5)에서 α는 허용하는 변이율을 나타내는 파라미터로서 작은 양의 실수값(예, 0.02)이다. 기호는 언급한 예와 같이 H, E, L 등 3개의 기호를 사용하는 것뿐만 아니라, 더 많은 기호를 사용하여 등급을 부여할 수도 있다. 단, 기호를 사용할 때는 타당한 기준을 선정하는 것이 중요하다. (그림 5)의 예에서는 직전값을 현재값과 비교하는 방식을 채택하고 있지만, 분석의 목적에 따라 최초 정상 샘플에 대한 정보에 기초한 값을 기준으로 삼을 수도 있고, 여러 가지 변형을 고려해 볼 수 있다.

(그림 7)은 제안한 코딩 방법에 따라 (그림 6)의 발현량을 나타낸 마이크로어레이 데이터에 대해 기호코딩한 결과를 나타낸 예이다. (그림 7)에서 마지막 열은 이전 열들의 기호를 결합하여 하나의 문자열로 표시한 것이다.

ID	A	B	C	D
Hs.487076-S	63.66716	64.75797	80.66199	96.12968
GI_28372504-S	335.9957	364.8326	396.2822	447.7596
GI_14702161-A	2209.12	2469.901	2829.83	3256.152
GI_4504606-S	1752.524	1871.283	2301.873	2434.09
GI_6912299-A	869.4264	1097.045	1257.373	1346.648
GI_11496268-S	2841.112	2560.534	2536.682	2243.37
GI_13259542-A	1454.968	1283.041	1342.546	1465.759
GI_16445418-I	424.3517	345.3897	441.4937	536.502
GI_16950622-A	97.72508	55.38776	75.00605	105.1253
GI_17105398-A	5769.672	5402.601	4750.161	4249.929
GI_18104974-I	787.7455	658.1235	750.8204	1212.208
GI_18104977-S	4243.022	3616.572	3195.556	3332.502
GI_18105008-S	2001.142	1931.459	2017.829	2164.687
GI_18379326-S	607.0491	512.0612	594.9178	679.1838
GI_21361740-S	3649.592	3535.912	3195.624	3254.379
GI_21361826-S	862.2292	799.6398	705.6665	812.965
GI_21450821-S	141.688	134.9264	129.7489	171.0047
GI_22208978-I	77.40584	62.61523	85.00197	87.83341

그림 6. 마이크로어레이 데이터의 예

ID	B	C	D	code
Hs.487076-S	E	H	H	EHH
GI_28372504-S	H	H	H	HHH
GI_14702161-A	H	H	H	HHH
GI_4504606-S	H	H	H	HHH
GI_6912299-A	H	H	H	HHH
GI_11496268-S	L	E	L	LEL
GI_13259542-A	L	H	H	LHH
GI_16445418-I	L	H	H	LHH
GI_16950622-A	L	H	H	LHH
GI_17105398-A	L	L	L	LLL
GI_18104974-I	L	H	H	LHH
GI_18104977-S	L	L	H	LLH
GI_18105008-S	L	H	H	LHH
GI_18379326-S	L	H	H	LHH
GI_21361740-S	L	L	E	LLE
GI_21361826-S	L	L	H	LLH
GI_21450821-S	L	L	H	LLH
GI_22208978-I	L	H	H	LHH

그림 7. 기호코딩한 마이크로어레이 데이터

3.5 관심패턴 후보 식별

각 유전자에 대해서 기호로 코딩된 문자열이 만들어지면, 분석자의 분석 목적에 따라 이에 대응하는 유전자를 검색할 수 있다. 예를 들어, LLH라는 기호문자열을 갖는 유전자를 찾으면, 이는 시간의 경과에 따라 발현량이 지속적으로 증가하는 것들을 찾은 것이 된다.

(그림 8)는 제안한 방법으로 LLH 패턴을 가진 유전자들을 찾고, 이들 유전자의 실제 발현 패턴을 도시한 예이다. 이 그림에서 보는 바와 같이 절대적인 실제값이 아니라 동일한 추이를 보이는 유전자를 효과적으로 찾을 수 있게 한다.

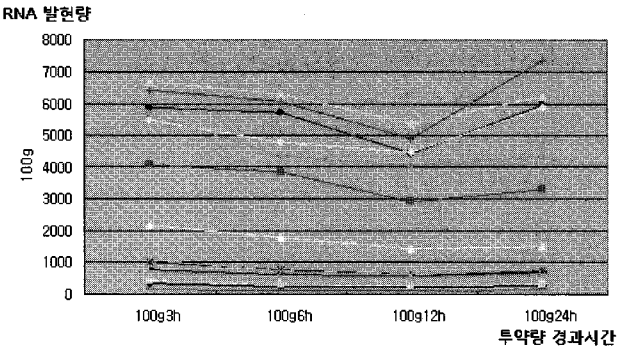


그림 8. 시간경과에 따라 LLH 추이를 보이는 유전자들의 발현패턴

4. 적용 예

제안된 기법의 유용성을 보이기 위해 실제 약물 반응 실험에 대한 마이크로어레이 데이터에 대해서 제안한 방법을 적용하였다. 실험에서는 생쥐(mouse)들에 대해서 약물을 전혀 투약하지 않은 집단, 50g 투약한 집단, 100g 투약한 집단으로 나누고, 약물 투입 3시간후, 6시간후, 12시간후, 24시간후에 샘플을 채취해서, 마이크로어레이를 통해서 유전자의 RNA 농도를 측정하였다. 실험에 사용된 마이크로어레이 칩은 Illumina사의 BeedChip[5]이 사용되었는데, 이 칩은 47,293개의 유전자 프로우브(probe)를 가지고 있기 때문에 이 만큼의 데이터값이 샘플별로 획득된다.

실험에서는 11개의 샘플에 대한 마이크로어레이 데이터가 제공되었고, 감지 p 값을 0.05 미만으로 설정하였을 때, 47,293개의 데이터 중에서 11,100개의 유전자가 선택되었다. 분석을 위한 데이터를 선정하기 위해서 약물 비투여 집단에서 시간의 경과에 따라 큰 차이를 보이지 않는 유전자를 선택하기위해 CTR_{var} 값을 0.05 미만으로 설정한 결과 11,100개 중에서 423개가 선택되었다. 이렇게 선택된 유전자가 분석 대상이 되며, 이들 유전자에 대해서 제안된 기호코딩 방법을 α를 0.02로 설정해서 적용하여 유전자들의 약물 투약량별, 시간경과에 따른 RNA 발현량에 대한 추이를 기호화된 패턴으로 나타냈다.

기호 코딩된 결과에 대해서, 추가 분석을 위한 유전자의 후보로는 약물 투약량에 무관하게 동일한 추이를 보이는 유전자를 선택하였다. 이 결과 47,293개의 유전자로 구성된 데이터로부터 약물반응 후보군으로 선택된 423개의 기호코딩된 유전자들 중에서 55개의 유전자가 일관성있는 추이를 보였다. 이러한 추이 패턴으로 EHH (1개), HHH(4개), HHL(6개), HLH(4개), HLL(9개), LEH(1개), LEL(1개), LHH(8개), LHL(1개), LLE(1개), LLH(15개), LLL(4개) 등이 나타났다.

추이 패턴을 추출함에 있어서 동일 투약량을 갖는 집단에 대해서 시간에 따른 추이를 보는 것과 함께, 약물 투약량에 따른 추이를 함께 고려하여 일정 패턴을 갖는 유전자를 추출하는 것도 유용한 정보를 제공할 수 있다. 이러한 정보를 획득하기 위해서는 제안된 기호코딩 방법을 동일 투약량을 갖는 집단의 시간에 따른 추이 패턴과 동일 시간대별 투약량의 증가에 따른 추이 패턴을 적용하면 된다. (그림 9)에서 가장 위쪽에 위치한 유전자는 약물투여량에 따라 발현량이 감소(코드값은 LL)하면서, 또한 시간이 경과하면서도 감소(코드값은 LLL)하는 추이를 보이는 것이다.

47,293 x 11 크기의 마이크로어레이 데이터에 적용하는 실험에서 제안한 방법에 따라 전처리와 후보 집단을 선정하고, 동일 투약량 샘플에 대한 시간의 경과에 따른 RNA 발현량 추이와, 투약량에 따른 시간대별 추이를 제안한 방법에 따라 코딩하여 데이터를 표현함으로써, 향후 추가분석 및 실험을 위한 유전자들에 대한 정보를 제공할 수 있었다.

5. 결 론

대용량, 고정밀도 마이크로어레이 기술의 발전은 유전자의 기능 예측, 생체네트워크 재구성 등에 많은 기여를 할 것이다. 마이크로어레이 실험에서 대량으로 생성되는 데이터를 효과적으로 분석하는 것은 주요 이슈로 부각되고 있다. 대상 유전자의 개수가 많기 때문에 의미있는 유의한 유전자를 자동으로 추출하는 기술의 확보가 필요하다. 특히, 약물효능 분석 등에 마이크로어레이가 활용되는 경우에는 데이터간의 관계를 고려한 데이터분석이 필요하다. 이 논문에서는 이러한

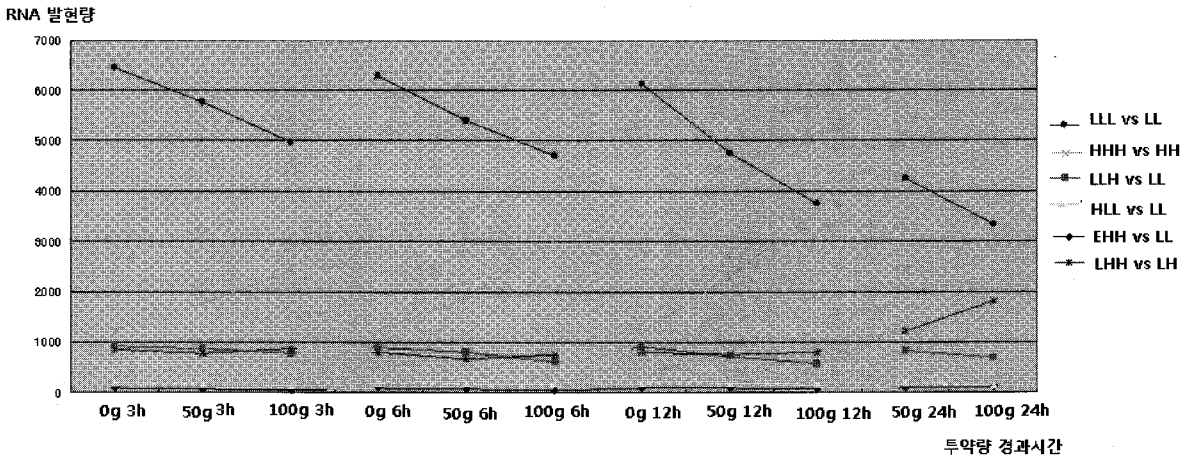


그림 7. 투약량과 경과시간에 따른 추이

분야에서 마이크로어레이 데이터를 효과적으로 분석할 수 있는 기법을 소개하였다.

제안한 분석 방법은 동일한 양의 약물을 투여한 후, 시간에 따른 추이에서 특정 패턴을 보이는 것을 검출하는 것과, 투여량의 증가에 패턴을 추출하는 것도 가능하다. 한편, 동일한 패턴을 보이는 유전자 집단으로 식별된 유전자들간에는 correlation이 있을 수 있기 때문에, 대사네트워크, 신호전달 네트워크, 조절네트워크에 대한 정보를 추가적으로 활용함으로써, 유전자들의 네트워크에서의 역할에 대한 추정뿐만 아니라 약물의 영향을 분석하는데도 유용하게 활용할 수 있을 것이다.

개발한 기법은 실제 약물의 동물실험 샘플에 대한 마이크로어레이 데이터에 대해서 적용하였으며, 제안한 기법에 의해 추천된 유전자의 약물에 대한 반응효과에 관한 분자생물학적인 추가적인 연구가 진행되고 있다. 이러한 적용을 통해서 제안된 기법이 약물 효능 분야의 마이크로어레이 데이터 분석에 유용함을 확인했다.

제안한 방법은 약물효능 분야의 마이크로어레이 데이터분석에서, 후보가 될 수 있는 유전자를 전처리 과정으로서 선정하는 방법과 발현량의 값을 기호로 변화하여 분석자가 기대하는 패턴을 보이는 유전자를 추천할 수 있도록 한다. 제안한 방법은 SOM, k-means 알고리즘을 적용하여 추출할 수 있는 패턴 군집보다 효과적으로 기대하는 추이 패턴을 보이는 집단을 추출할 수 있는 마이크로어레이 데이터 분석 등 추이패턴 추출이 필요한 분야에서 유용한 방법이다.

참 고 문 헌

[1] D. W. Mount, *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Habor Lab Press, 2004.

[2] G. B. Forgel, D. W. Corne, *Evolutionary Computation in Bioinformatics*, Morgan Kaufmann Publishers, 2003

[3] E. Alpaydin, *Introduction to Machine Learning*, MIT, 2004.

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

[5] Illumina, *BeadStudio Gene Expression Module User Guide*, Illumina, 2006.

[6] R. O, Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, John Wiley & Sons, INC., 2001.

[7] So. Draghici, *Data Analysis Tools for DNA Microarrays*, Chapman & Hall/CRC, 2003.

저 자 소 개

이선아
15권 3호 참조

이건명
15권 3호 참조

김원재
충북대학교 의과대학 교수