마이크로어레이 데이터의 기호코딩을 통한 유의한 후보 유전자 검출

Candidate Significant Gene Recommendation with Symbolic Encoding of Microarray Data

이건명[1], 이혜리[1], 김원재[2], 윤석중[2], 김용준[2], 정필두[2], 김은정[2]

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

요 약*

마이크로어레이는 생명과학 분야에서 사용되는 대규모의 유전자 발현정도를 동시에 측정할 수 있는 도구이다. 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 시간의 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. 제안한 방법은 실제 약물 투여 실험 샘플에 대한 마이크로어레이 데이터에 적용하여 활용가능성을 확인하였다.

Key Words : bioinformatics, microarray, data analysis, data mining

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. 마이클로어레이(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 투약

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

량을 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 4만여종 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.

2. 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 약물개발에서 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.

3. 제안한 기호코딩 기반의 약물에 반응하는 유의한 유전자 추출 기법

마이크로어레이 데이터는 (그림 1)과 같이 많은 수의 유전자에 대한 발현량에 대한 수치값 및 관련정보로 구성된다. 이 연구가 수행되는 시점에서 Illumina사에서는 한번에 최대 47,000여 유전자의 발현량을 측정할 수 있는 마이크로어레이를 제공 중이었다. 약물에 유의한 반응을 보이는 후보를 추출하기 위해서는 가능하면 전처리과정에서 연관성이 없을 것으로 판단되는 많은 유전자를 고려대상에서 제외하는 것이 필요하다. 이를 위해서는 약물 효능 실험에 대한 특징과 분석배경지식을 활용하는 것이 필요하다.

TargetID	A	B	C	D
Hs.487076-S	89.50107	63.66716	55.11445	87.91047
GI_28372504-S	373.2008	335.9957	262.6891	375.5819
GI_14702161-A	2235.341	2209.12	2269.169	2169.262
GI_4504606-S	1244.967	1752.524	2420.781	1240.191
GI_6912299-A	971.8798	869.4264	1060.892	954.3829
GI_6912389-S	316.0078	297.7956	316.4807	318.613
GI_38044109-S	354.9342	370.5561	365.3142	347.2064
GI_31543135-S	267.9129	263.1059	263.0451	267.7958
hmm30087-S	61.87408	55.27253	56.3293	62.68459
GI_31543651-S	2650.234	2252.034	2333.348	2782.082
GI_41152082-S	1078.54	1029.648	1009.283	1067.719
GI_40804466-S	2796.469	2676.048	2820.671	2784.323
GI_40538727-S	422.1987	336.0438	295.1317	410.5634
GI_24308302-S	409.2253	364.173	362.6505	423.3194
GI_26553431-S	718.9124	644.465	687.3035	752.6283
GI_4507106-S	1767.297	1507.52	1460.917	1761.637
GI_7661555-S	1175.305	1097.245	1180.593	1186.275
GI_31542527-S	1322.035	1385.067	1313.93	1320.688

그림 1. 마이크로어레이 데이터의 예

제안한 방법에서는 약물반응에 유의한 후보유전자 선정을 위해서 다음과 같은 기법을 사용한다. 우선, 마이크로어레이 실험에서 측정치에 대한 신뢰도가 떨어지는 것을 우선적으로 제외한다. 마이크로어레이 스캔 도구들은 마이크로어레이에서 각 유전자에 대응하는 위치로부터 발현정도에 대한 값을 읽어들일 때 이미지형태로 읽으면서 측정치에 대한 정확도를 detection p-value라는 값을 제공하는데, 이 값이 특정값(예, 0.05) 이상이면 추후 분석에서

고려할 유전자 집단에서 배제한다. 분자생물학 차원에서 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 우선 추천한다. 이러한 유전자를 선택하기 위해서, 먼저 정상 집단 샘플에 대해서 각 유전자별로 평균 발현량 CTR_{avr}을 계산한다. 마이크로어레이 분석에서는 특정 유전자의 발현량이 비정상으로 큰 값을 갖는 경우도 있기 때문에 값이 커지는 부분에 대해서는 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, 정상 집단에 대해서 각 유전자 별로 최대 발현량과 최소 발현량의 차이 CRT_{int}를 계산한다. 변화량이 적은 유전자들을 선택하기 위해서, 유전자별로 발현량의 변화율을 $CTR_{var} = CTR_{int}/CTR_{avg}$를 이용하여 계산한 다음, 이를 오름차순의 정렬한다. 정렬된 유전자 리스트에서 CTR_{var}값이 일정값 (예, 0.05)이하인 것들을 추후 고려할 유전자 집단으로 선정한다. 이와 같은 과정을 거치게 되면, 고려할 유전자가 수가 수만종에서 수백종까지 줄일 수 있게 된다.

약물효능을 분자생물학적 수준에서 유전자의 발현패턴을 보고 추정하려고 하는 연구에서는, 약물의 투여 및 투여량에 따라 특정한 패턴을 보이는 유전자를 식별해 내고, 이들에 대해서 추가적인 연구를 하게 된다. 마이크로어레이 데이터는 수치로 되어 있기 때문에, 이들 수치값으로 많은 수의 유전자들로 일정 패턴을 갖는 것을 찾는 것은 쉽지 않다. 제안한 방법에서는 발현정도를 나타내는 수치값을 기호로 변환한 다음 특정 패턴을 보이는 유전자를 추출하도록 한다. 이를 위해서 정상 샘플들과 약물투여 집단의 샘플들에 대해서, 우선 각 유전자별로 평균 발현정도 ALL_{avg}를 결정한다. 이 평균 발현정도값을 기준으로 직전 값($preceed_{val}$) 대비 현재 값(cur_{val})을 보고 현재값의 상태를 나타낸다. 예를 들면 이전값보다 현재값이 일정수준이상 크면, 현재값을 'H'로 변경하여 코딩하고, 비슷하면 'E'로, 일정수준보다 작으면 'L'로 변경하고 코딩한다. 다음은 위의 H, E, L 기호를 사용하는 코딩 방법을 예시한 것이다.

if (abs($preceed_{val}$-cur_{val}) < ALL_{avg}*α), 'E'로 코딩

if ($preceed_{val}$ > cur_{val}+ALL_{avg}*α), 'L'로 코딩

if ($preceed_{val}$ < cur_{val}+ALL_{avg}*α), 'H'로 코딩

그림 2. 코딩 규칙의 예

(그림 2)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 2)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.

TargetID	B	C	D	E	Code String
GI_28395048-S	H	L	L	H	HLLH
GI_18104977-S	H	L	L	H	HLLH
Hs.487076-S	L	E	H	H	LEHH
GI_38044109-S	H	H	H	L	HHHL
GI_4507106-S	L	H	L	L	LHLL
GI_34932413-S	H	L	L	H	HLLH
GI_28372504-S	L	H	H	H	LHHH
hmm8085-S	L	L	H	H	LLHH
GI_13259542-A	H	L	H	H	HLHH
GI_40538727-S	L	H	L	H	LHLH
GI_37552553-S	H	L	L	L	HLLL
GI_30795211-S	H	L	L	L	HLLL
GI_14702161-A	E	H	H	H	EHHH
GI_27777635-S	L	L	L	H	LLLH
GI_4504864-S	H	L	L	H	HLLH

그림 3. 기호코딩한 마이크로어레이 데이터

(그림 3)은 제안한 코딩 방법에 따라 발현량의 수치값을 기호로 나타낸 예이다. 그림에서 마지막 행은 이전 행들의 기호를 결합하여 하나의 문자열로 표시한 것이다. 이와 같은 과정을 통해서 각 유전자에 대해서 기호들의 문자열이 만들어지면, 분석자의 분석 목적에 따라 이에 대응하는 유전자를 검색할 수 있다. 예를 들어, HHHH라는 기호문자열을 갖는 유전자를 찾으면, 이는 시간의 경과에 따라 발현량이 지속적으로 증가하는 것들을 찾은 것이 된다.

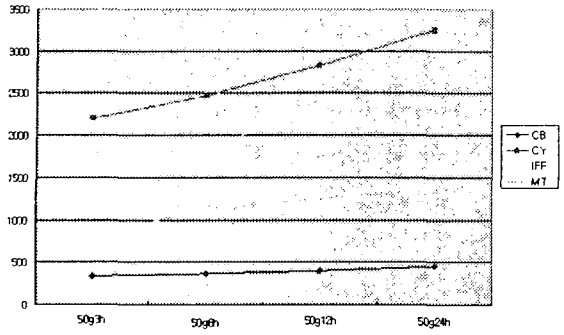

그림 4. HHHH 패턴을 갖는 유전자의 실제 발현량의 패턴 예

(그림 4)는 제안한 방법으로 HHHH 패턴을 가진 유전자들을 찾고, 이들 유전자의 실제 발현 패턴을 보인 예이다. 이 그림에서 보는 바와 같이 절대적인 실제값이 아니라 동일한 추이를 보이는 유전자를 효과적으로 찾을 수 있게 한다.

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, 신호전달네트워크, 조절네트워크에 대한 정보를 추가적으로 활용함으로써, 유전자들의 네트워크에서의 역할에 대한 추정뿐만 아니라 약물의 영향을 분석하는데도 유용하게 활용할 수 있을 것이다.

4. 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 분야에서 마이크로어레이 데이터를 효과적으로 분석할 수 있는 기법을 소개하였다. 제안한 방법에서는 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 기법을 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 약물효능 분야의 마이크로어레이 데이터 분석에 유용함을 확인했다.

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.