3차원 유전자 발현 데이터에서의 시간 관계 규칙을 이용한 유전자 상호작용 조절 네트워크 구축[1)]

Meijing Li, 박진형, 이헌규, 류근호
충북대학교 데이터베이스/바이오인포메틱스 연구실
e-mail:{mjlee, neozean, hglee, khryu}@dblab.chungbuk.ac.kr

Constructing Gene Regulatory Networks using Temporal Relation Rules from 3-Dimensional Gene Expression Data

Meijing Li, Jin Hyoung Park, Heon Gyu Lee, Keun Ho Ryu
Database/Bioinformatics Laboratory, Chungbuk National University

요 약

유전자들은 복잡한 상호작용을 통해 세포의 기능이 조절된다. 상호작용하는 유전자 그룹들을 유전자 조절 네트워크라고 한다. 기존의 유전자 조절 네트워크는 2D microarray 데이터를 이용하여 시간의 흐름에 따른 유전자간의 상호작용을 알 수가 없었다. 이 논문에서는 시간의 변화에 따른 유전자들 간의 조절관계를 살펴 볼 수 있는 조절네트워크 모델링의 방법을 제시한다. 유전자의 발현양을 표시하기 위해 이진 이산화 방법을 사용하였고 3D microarray 데이터에서 유전자 발현 패턴을 찾기 위해 Cube mining 알고리즘을 적용하였고, 유전자간의 관계를 밝히기 위해 시간 관계 규칙탐사 기법을 사용하여 유전자들 간의 시간 관계를 포함한 유전자 조절네트워크를 구축하였다. 이 연구는 시간의 흐름에 따른 유전자간의 상호작용을 알 수 있으며, 모델링된 조절 네트워크를 이용하여 기능이 아직 발견되지 않은 유전자들의 기능을 예측 할 수 있다.

1. 서론

유전자들이 발현하는 데는 여러 가지 환경적인 요소가 많이 작용하는데 그 요소들 중 하나가 그 유전자의 발현에 영향을 미치는 다른 유전자들과의 관계이다. 유전자는 상호 복합적으로 작용을 하기 때문에 유전자들의 조절 관계를 분석하는 것은 유전자의 기능을 예측하는 데에도 필수적이다. 최근 DNA microarray 와 같이 유전자 발현 기술은 유전자 레벨에서 유전자 발현 패턴을 관찰할 수 있도록 해주었으며 이를 기반으로 하여 유전자들 간의 상호 조절 관계를 분석할 수 있게 되었다.

유전자들의 그룹은 복잡한 상호작용들을 통해 세포의 기능이 조절되며 이러한 상호작용을 하는 유전자 그룹들을 유전자 조절 네트워크(GRNs: gene regulatory networks)라고 한다.

기존에 이러한 조절관계를 밝히는 연구[4],[5],[6]는 유전자간의 상호조절관계를 밝히기 위해 2D microarray 데이터에서 발현 양을 이용하여 사용자가 정의 한 지지도와 신뢰도를 이용한 빈발 패턴 마이닝을 적용하여 빈발한 유전자들을 추출한다. 이렇게 추출된 빈발한 유전자들 간의 조절자를 예측하기 위해 조건부 확률 또는 베이지안 네트워크등과 같은 방법을 이용하여 조절자를 예측한다. 예측된 조절자를 연결하여 유전자 조절 네트워크를 모델링한다. 하지만, 유전자 간의 상호조절관계는 시간의 흐름은 나타나지 않고 단지 특정시점에서의 유전자 간의 상호 조절관계를 알아 볼 수 있다.

이 논문에서는 시간의 변화에 따른 유전자간의 상호조절관계를 관찰하고 유전자 기능을 예측하기 위해 3차원 microarray 데이터 Time-Sample-Gene으로부터 유전자 조절 네트워크를 모델링 한다. 조절네트워크를 모델링하기 위해 3D microarray로 부터 직접 유전자 발현 패턴을 찾을 수 있는 3D cube mining을 이용하여 유전자 발현 패턴을 찾는다. 다음으로 시간의 흐름에 따른 유전자간의 상호조절관계를 관찰하기 위해 유전자간의 시간 간격을 시간 연산자와 비교하여 조절 네트워크를 구축한다.

논문의 구성은 다음과 같다. 2장에서는 유전자들간의 조절관계를 찾고자 하는 3D 유전자발현 데이터의 이산화 과정을 기술하고, 3장에서는 cube mining closed 패턴 탐사 알고리즘 수행 과정에 대해 서술한다. 4장에서는 생성된 유전자 발현 패턴들로 시간 관계 연산자를 이용하여 유전자들의 시간 관계를 찾고 시간 관계에 따라 조절 네트워크를 구축하는 방법과 절차에 대해 기술한다. 5장에서는 논문에서 제안한 방법에 대한 성능평가 실험에 대해

1) 이 논문은 한국과학재단에서 지원하는 우수연구 센터사업 종양의료 개인특화를 위한 기기, 시스템 연구센터(ERC)의 2008년도 연구과제 지원에 의한 결과임.

서술하고, 6장에서는 전체적인 논문에 대해 결론을 맺는다.

2. 유전자 발현 microarray 데이터 전처리

이 단계에서는 closed 패턴을 찾는 cube mining 알고리즘에 적용할 수 있도록 유전자 발현 원시적인 데이터에 대해 이산화 전처리를 한다. 3차원 유전자 발현 데이터샘플은 (그림 1)과 같다.

• T=10min

gene/sample	sample1	sample2	sample3	sample4	sample5
YAL004W	342	305	37	247	207
YAL005C	3979	296	3683	1370	227
YAL007C	342	305	37	247	207
YAL008W	401	304	97	333	218

• T=30min

gene/sample	sample1	sample2	sample3	sample4	sample5
YAL004W	348	312	36	249	213
YAL005C	7430	355	7075	1108	235
YAL007C	336	298	38	252	217
YAL008W	508	338	170	397	219

• T=50min

gene/sample	sample1	sample2	sample3	sample4	sample5
YAL004W	360	329	31	298	260
YAL005C	5932	352	5580	1076	263
YAL007C	360	329	31	298	260
YAL008W	466	335	131	432	277

(그림 1) 3D gene microarray data

유전자 발현 데이터는 유전자의 발현 여부만 확정 할 수 있으면 된다. 즉 원본 데이터가 이산화 과정을 거친 후 발현 비율이 높은 유전자를 1로, 발현 비율이 낮은 유전자를 0으로 나타낸다. Time-sample-gene 데이터 집합을 $O'=T\times S\times G=\{O'_{k,i,j}\}$ ($k\in[1, l]$, $i\in[1, n]$, $j\in[1, m]$)라고 정의할 때 식(1)과 같은 방법으로 데이터 이산화를 한다.[1]

$$O_{k,i,j}=\begin{cases}1 & O'_{k,i,j}\geq\dfrac{\sum_{j=1}^{m}O'_{k,i,j}}{m}\text{ 일 때,}\\ 0 & O'_{k,i,j}>\dfrac{\sum_{j=1}^{m}O'_{k,i,j}}{m}\text{ 일 때,}\end{cases}\qquad \text{식(1)}$$

이산화 한 다음의 유전자 발현 데이터는 (그림 2)와 같다.

3. Cube mining 알고리즘을 적용한 3차원 유전자 발현 closed 패턴 생성

이 단계에서는 2장의 전처리과정에서 이산화한 데이터로 cube mining 알고리즘을 적용하여 time-sample-gene 유전자 발현 closed 패턴을 찾는다[1]. cube mining 알고리즘을 수행하여 얻은 패턴은 "$time_2$ $time_3$ $time_7$ $time_9$: $sample_2$ $sample_4$ $sample_8$: $gene_1$ $gene_2$ $gene_4$ = 4:3:3"의 형태로 출력되는데 시간이 $time_2$, $time_3$, $time_7$, $time_9$일 때, $sample_2$, $sample_4$, $sample_8$에서 유전자 $gene_1$, $gene_2$, $gene_4$이 동시에 발현 한다는 것을 의미하고, 4:3:3은 time, sample, gene 각각의 support 값을 나타낸다.

T=10min

gene/sample	sample1	sample2	sample3	sample4	sample5
YAL004W	0	1	0	0	0
YAL005C	1	0	1	1	1
YAL007C	0	1	0	0	0
YAL008W	0	1	0	0	1

T=30min

gene/sample	sample1	sample2	sample3	sample4	sample5
YAL004W	0	0	0	0	0
YAL005C	1	1	1	1	1
YAL007C	0	0	0	0	0
YAL008W	0	1	0	0	0

T=50min

gene/sample	sample1	sample2	sample3	sample4	sample5
YAL004W	0	0	0	0	0
YAL005C	1	1	1	1	0
YAL007C	0	0	0	0	0
YAL008W	0	0	0	0	1

(그림 2) 이산화 한 후 3D gene microarray data

각각의 최소지지도 임계값은 시간 최소지지도 minT=2, 샘플에 대한 최소지지도 minS=2, 유전자 최소지지도 minG=1로 설정하고, 위의 샘플 데이터에 대해 Cube mining 알고리즘을 수행하면 아래와 같은 3개 closed 패턴을 얻게 된다. (t_2 t_3 : g_2 : s_1 s_2 s_3 s_4, 2:1:4), (t_1 t_2 : g_2 : s_1 s_3 s_4 s_5, 2:1:4), (t_1 t_2 t_3 : g_2 : s_1 s_3 s_4, 3:1:3). 여기서 t는 "time"을 나타내고 g는 "gene"을 나타내며 s는 "sample"을 나타낸다.

4. 시간 관계 규칙 탐사 기법을 사용한 유전자 발현 상호작용 조절네트워크 구축

Cube mining 알고리즘을 수행하여 얻은 closed 패턴들로 시간 연산자를 사용하여 유전자들간의 시간 관계[3]를 찾고, 그 시간 관계로부터 유전자 발현 상호작용 조절네트워크를 구축한다.

Cube mining 알고리즘을 수행하여 얻은 3차원 데이터 패턴을 1:1:1의 형태로 고쳐준다. 예를 들면 패턴 (t_2 t_3 : g_2 : s_1 s_2 s_3 s_4, 2:1:4)를 1:1:1의 형태로 고쳐주면 (t_2:g_2:s_1), (t_2:g_2:s_2), (t_2:g_2:s_3), (t_2:g_2:s_4), (t_3:g_2:s_1), (t_3:g_2:s_2), (t_3:g_2:s_3), (t_3:g_2:s_4) 등 8개의 패턴을 얻게 된다. 모든 cube mining 알고리즘 결과 패턴을 1:1:1의 형태로 분리하면 중복된 패턴들이 나타나게 된다. 이런 중복 패턴들은 모두 제거하고, 패턴들에 나타났던 유전자들을 빈발 유전자 종류 집합이라고 하며, 예제 (그림 3, 좌)와 같이 정렬된 데이터베이

스를 얻게 된다.

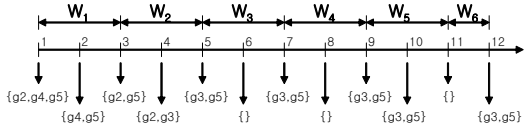
Sample	Time	Gene
s_1	t_1	g_2,g_4
s_1	t_2	g_4,g_5
s_1	t_3	g_5,g_2
s_1	t_5	g_3
s_1	t_7	g_5,g_3
s_2	t_1	g_2,g_5
s_2	t_4	g_2
s_2	t_6	g_5,g_3
s_2	t_9	g_5,g_3
s_3	t_1	g_2,g_4
s_3	t_3	g_2,g_5
s_3	t_4	g_2
s_3	t_7	g_5,g_3
s_3	t_{10}	g_5,g_3
s_4	t_1	g_4
s_4	t_4	g_3
s_4	t_5	g_5,g_3
s_4	t_{12}	g_5,g_3

Sample	$Time_{vs}$	$Time_{ve}$	Gene
s_1	t_1	t_3	g_2
s_1	t_5	t_7	g_3
s_1	t_2	t_7	g_5
s_2	t_1	t_4	g_2
s_2	t_6	t_9	g_3
s_2	t_1	t_9	g_5
s_3	t_1	t_4	g_2
s_3	t_7	t_{10}	g_3
s_3	t_3	t_{10}	g_5
s_4	t_4	t_{12}	g_3
s_4	t_5	t_{12}	g_5

(그림 3) 정렬된 데이터베이스(좌)와 일반화된 시간 간격 데이터베이스(우)

발생 시점만 가지는 데이터를 시간 간격을 가지는 데이터로 변환하기 위하여 유전자의 발현 전반 시간 과정에서 특정 시간대에만 집중적으로 발현하는 유전자들을 가지치기하고, 지속적으로 균등하게 발현하는 유전자들만을 일반화 한다. 이런 균등 유전자 종류 집합[3]을 찾기 위해 시간 윈도우 크기 W_{size}와 유전자 빈발도 $Freq_{min}$라는 두 개의 파라미터를 사용한다.

최초의 빈발 유전자가 발생한 시간부터 마지막 유전자가 발생한 시간까지의 전체 기간을 처음 시작점부터 사용자가 지정한 윈도우 크기 W_{size}에 따라 나누고 W_i 안에 빈발 유전자 종류 G_i가 있으면 G_i는 W_i에서 발생한다고 한다. 이때 유전자 빈발도 $Freq(G_i)$는 윈도우 시퀀스 $W_1W_2W_3\cdots W_n$에서 G_i가 발생한 W_i개수를 말한다. 사용자의 요구에 따라 윈도우 크기와 $Freq_{min}$값을 주고, $Freq_{min}$을 만족시키지 못하는 빈발 유전자 종류를 제거하면 균등 유전자 종류 집합을 얻게 된다(그림 4). 예제에서 W_{size}=2, $Freq_{min}$=33%로 하고 얻은 균등 유전자 종류 집합은 $\{g_2,g_3,g_5\}$이다.



(그림 4) 시간 간격을 가진 유전자 시퀀스 예제

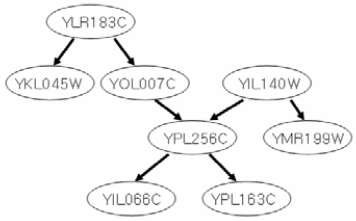
매개 Sample에서 나타난 균등 유전자 종류들을 $<G_i,T_s,T_e>$의 형태로 시간 일반화를 한다. 여기서 유전자 종류 G_i가 샘플 s_i에서의 시작 시점을 T_{vs}로 표시하고 마지막 시점을 T_{ve}로 표시한다.

일반화된 시간 간격 데이터베이스는 (그림 3, 우)과 같은데 여기에서 각각의 샘플에서 균등 유전자 종류 G_i들간의 시간 관계 규칙들 before(G_i,G_j), during(G_i,G_j), overlap(G_i,G_j)을 찾아낸다. 이 세 가지 시간 관계들에서 시간 관계 지지도가 높을수록 먼저 발현하는 유전자가 나중에 발현한 유전자의 조절자일 가능성이 높다. [3]에서의 후보시간 관계 트리로 부터 사용자가 지정한 시간 관계 최소지지도 $Supp_{min}$이상을 가지는 빈발 시간 관계 집합 $FR=\{R_2(x,y),\cdots R_z(x,y)\}$을 구한다. 예를 들어 유전자 g_2와 g_4의 시간 관계인 before(g_2,g_4)의 지지도가 80%일 경우, 유전자 g_2가 유전자 g_4보다 먼저 발현 되는 비율이 80%라는 것을 의미하며, 유전자 g_2가 유전자 g_4의 발현 조절자이다(조절관계 표현:$g_2\rightarrow g_4$). (그림 3, 좌)의 예제로부터 위와 같은 방법으로 찾은 빈발 시간 관계와 그 시간 관계들로부터 얻은 유전자 조절 네트워크는 (그림 5)과 같다.

(그림 5) 빈발 시간 관계로부터 조절 네트워크 구축의 예

5. 실험 및 평가

Yeast cell-cycle 조절 유전자들에 대한 CDC15실험[10] 데이터를 이 논문의 실험데이터로 하였는데 10분 간격으로 9개의 샘플에서의 7,761개의 유전자 ORF 발현 상황에 대해 측정한 결과를 보여주며 실험 70분 후부터 250분까지 19개의 시점의 데이터를 실험에 사용하였다. 실험에 사용된 유전자 데이터의 조절네트워크는 (그림 6)과 같다.



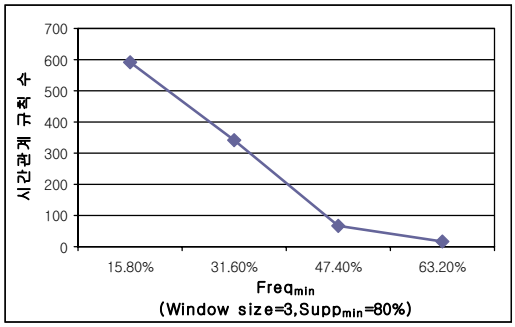
(그림 6) 유전자 조절 네트워크

<표 1> 실험에서 사용되는 파라미터 변수

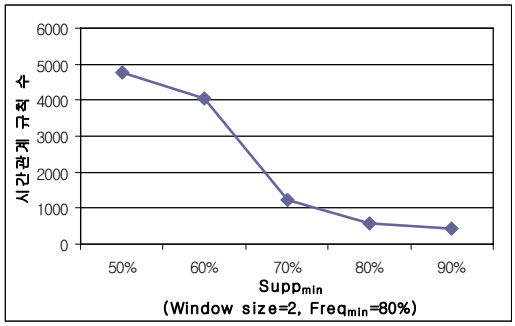
파라미터 변수	설명
W_{size}	시간 관계 규칙 탐사 기법에서의 윈도우 크기
$Freq_{min}$	전반적인 시간 축에서의 유전자 최소 빈발도
$Supp_{min}$	시간 규칙 최소 지지도

이 논문에서 제안한 방법의 시간 관계 규칙 탐사 기법의 유전자 시간 관계를 찾는 데 대한 실험을 수행하였다. 시간 관계 규칙 발견에 필요한 파라미터들은 <표 1>과 같고 각 파라미터 변화에 대한 생성되는 규칙의 개수에 대

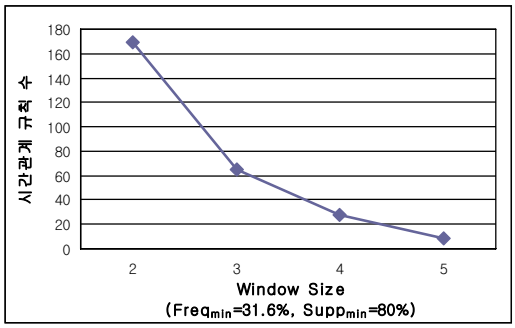
한 실험 결과는 (그림7)과 같다.



(그림 7-a) 최소 빈발도에 따른 생성된 시간 관계 규칙 수



(그림 7-b) 최소 지지도에 따른 생성된 시간 관계 규칙 수



(그림 7-c) 윈도우 크기에 따른 생성된 시간 관계 규칙 수

(그림 7)에서 보여주는 것처럼 윈도우 크기와 $Freq_{min}$ 값은 크게 줄수록 생성되는 시간 관계 수가 모두 적어지지만 $Freq_{min}$ 값이 윈도우 크기보다 생성되는 유전자들의 시간 관계 규칙 수에 주는 영향이 현저히 더 크다는 것을 알 수 있다. $Supp_{min}$ 값이 생성된 시간 관계 규칙 수에 대한 영향을 보면 세 개의 파라미터들 중에서 가장 크며 60%에서부터 70%에서 급격히 내려가는 것을 볼 수 있는데 이것은 생성된 유전자들의 시간 관계들의 시간 관계규칙 지지도가 60%부터 70%에 많다는 것을 알 수 있다.

6. 결론

이 논문에서는 유전자 상호간 시간의 변화에 따른 조절 관계를 찾아내기 위해서 시간 연산자와 3차원 Cube mining을 이용한 시간 유전자 조절 네트워크 모델링 기법을 제안하였다.

제안된 유전자 조절 네트워크의 구축 과정은 첫째, 3D Cube mining 알고리즘을 적용하여 전처리 된 각 유전자 발현 데이터에서 빈발한 유전자 패턴들의 발견하였고 두 번째 단계에서는 유전자 패턴들로부터 유전자들의 시간 관계 규칙을 찾고 시간 연산자를 적용하여 네트워크를 구성하였다.

참고문헌

[1] Liping Ji, Kian-Lee and Anthony K. H. Tung “Mining frequent closed cubes in 3D datasets”, VLDB, Nov. 2006.

[2] Paul T. sellman, Gavin Sherlock, Michael Q.zhang, "comprehensive Identification of Cell Cycle-requlated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization". Molecular Biology of the Cell. 1998.

[3] 이용준, 서성보, 류근호,”시간간격을 고려한 시간관계 규칙 탐사 기법” 컴퓨터연구정보센터,2001.

[4] Friedman, N., Linial, M., Nachman, I. and Pe’er, D., "Using Bayesian networks to analyze expression data," Journal of Computational Biology, pp. 601-620, Apr. 2000.

[5] Dirk Husmeier, "Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks," Bioinformatics, pp. 2271-2282, Nov. 2003.

[6] Holter, N. S., Maritan, A., Fedoroff, N. V. and Banavar, J. R., "Dynamic modeling of gene expression data,” Proc., Natl, Acad. Sci., pp. 1693-1698, Feb. 2000.