은닉변수학습 모형에 기반한 시간적 프로파일을 이용한 조절 유전자군의 탐색

양진산° 장병탁
서울대학교 컴퓨터공학부 바이오지능연구실 및 바이오정보기술연구소(CBIT)
{jsyang, btzhang}@bi.snu.ac.kr

Searching for the regulated gene groups through temporal profiling of microarray expressions based on the latent variable learning model

Jinsan Yang°, Byoung-Tak Zhang
School of Computer Science and Engineering and Center for Bioinformation Technology (CBIT)

요 약

유전자 발현에 있어서의 조절작용은 유전자간의 복합적인 상호작용의 결과에 기인한다. 따라서 이러한 현상으로부터 기능적으로 연관된 유전자 군을 식별하기 위해서는 단일 유전자보다는 복수의 유전자군의 발현패턴을 대상으로 하게 된다. 이 경우 발현패턴의 시간에따른 다양하고 복잡한 특징들은 은닉변수학습 모형을 이용하므로서 보다 명확하게 표현될 수 있고, 유사한 기능을 가진 유전자 군을 탐색 하는데에 효과적으로 이용될 수 있다. 본 논문에서 제시된 은닉변수학습 모형은 이스트 cell cycle 데이터에 적용한 결과 특정 조절유전자에 대하여 생물학적으로 연관된 유전자 군을 찾는 데에 다른 방법과 비교하여 효과적임을 보일 수 있었다.

1. 서 론

생물학적인 현상의 연구에서 DNA 마이크로 어레이 기술은 동시에 다수 유전자의 발현 값을 측정 하므로서 유전자발현에 따른 다량의 정보를 효과적으로 얻게 하여준다. 대부분의 경우 하나의 유전자의 발현상태는 다른 유전자의 발현에 영향을 미치게 되고 이러한 관계에 있는 유전자 군을 예측하는 것은 그에 따른 생물학적 과정의 이해에 매우 유용하다.

여러 개의 유전자군들 중에서 기능적, 생물학적으로 연관 되어진 유전자 군을 탐색하기 위해서는 개별유전자의 발현 패턴보다는 각 유전자 군에 속한 유전자전체의 상호작용의 결과에 따른 발현특색들을 이용하여 분석하는 것이 효과적 이다. 즉 개별유전자의 발현은 다른 유전자와의 상호 조절작용의 결과물로 볼수 있기 때문에 개개의 유전자를 대상으로 하기보다는 조절작용에 관여하는 유전자 모두를 그 대상으로 하는 것이 보다 적절하게 된다.

이와 관련하여 단일 유전자 이외의 발현에 기반한 관련 연구로는, 서로 반대되는 발현 추세를 보이는 유전자의 기능적 분석을 위한 음으로 관련된 유전자발현의 사용 [2], 시간 천이 및 반전된 유전자 발현의 경우 유전자간의 다양한 생물학적 상관관계의 분석을 위한 다이나믹 프로그래밍에 기반한 방법 [3] 등이 알려져 있다. 두 방법은 모두 두 유전자간의 발현을 비교하고 있지만 대칭적인 발현관계[2]나 동일한 유전자에 대하여 다른 실험조건에 대한 유사도[3] 등에 대해서만 제한적으로 적용하고 있다.

본 논문에서는 조절유전자군에 속하는 유전자들 간의 일반적 발현 특징들을 정의하고 그변화에 따른 특성들을 시간에 따라서 재구성 하므로써 조절작용에 관여하는 유전자군들을 예측 하고저 한다. 이를 위한 방법으로 은닉변수학습 모형을 이용하여 많은수의 발현 특성들을 분석이 용이한 은닉변수로 줄여주고 이를 다시 시간에 따른 프로필로 나타내주는 과정을 거쳐서 유사한 조절유전자군의 예측에 이용하게 된다.

본 논문에서 사용된 데이터는 Spellman [4]의 yeast cell cycle data를 사용하였고 적절한 은닉변수학습 모형을 사용한 결과 기능적으로 유사한 유전자군들을 예측할 수 있었다. 또한 다른 방법으로 선택된 유전자군과의 비교에서 제시된 모형의 효과를 확인할수 있었다.

2. 이론적 배경

2.1 은닉변수학습 모형

은닉변수학습 모형은 관찰된 변수에 대하여 은닉변수를 추가로 가정하여 두 변수들 간의 관계를 설정한 모형이다. 따라서 은닉변수학습 모형을 이용하면 원래의 변수가 가지는 복잡한 분포를 좀더 다루기 쉬운 은닉변수의 분포로 바꾸어 주게 된다. 이때 원래의 변수가 가지는 복잡한 상관관계를 은닉변수을 이용하여 조건부 독립의 형태로 표현하므로서 계산하기 간편한 형태로 다시 표현할 수 있게 된다.

즉 관찰된 변수의 수를 D 라 하면 관찰된 변수 y 와 은닉변수 x 의 결합확률은 다음과 같이 나타내어 진다.

$$p(y,x) = p(x)p(y|x) = p(x)\prod_{i=1}^{D} p(y_i|x)$$

또한 조건부 확률식 $p(y|x)$ 에서 y 는 오차 e 를 수반한 x 의 사상 h 로 표시된다.

$$y = h(x,W) + e$$

그러므로 y 의 분포는 결합 확률 $p(y,x)$의 주변 확률로 주어지게 되고 은닉변수학습 모형은 $p(x)$, $h(x,W)$ 그리고 오차항 e 의 분포를 명시하므로서 결정된다.

2.2 은닉 격자 모형 (SOLL)

은닉격자 모형[5] 은 은닉변수학습 모형에서 $p(x)$의 분포를 은닉공간상에서 K 개의 노드를 가진 은닉 격자로 두고 은닉변수의 사상을 다음과 같이 가정한다.

$$y = \phi(x)W + e$$

여기서 $\phi(x)$는 M개의 기저함수에 대한 은닉변수 x의 사상이고 W 는 파라미터 행렬, e는 분산이 대각행렬로 주어지는 오차항이다.

은닉격자 모형은 y의 값을 입력으로 하여 K개의 노드에 대한 연결가중치를 구하고 이 연결가중치에 의한 은닉노드 값의 추정과 여기에서 얻어진 은닉변수의 값을 이용한 W 값의 추정을 반복하므로서 수행된다 (표1).

3. 분석방법

3.1 특성(Feature) 들 의 선택

잠재적 조절유전자군의 분석을 위해서는 그 무리에 속한 유전자간의 상호작용을 반영하는 특성들을 선택하는 것이 필요하다. 중요한 특성들로서는 유전자간의 시간에 따른 발현값의 차이, 전후의 발현 변화율의 차이, 국지적 상관계수, 곡률의 변화량 등이 사용 되었다.

(발현값의 변화) $|a_i - b_i|$

(기울기의 변화) $|(a_i - a_{i-1}) - (b_i - b_{i-1})|$

(국지적 상관계수) $|a_{i-1}b_{i-1} + a_i b_i + a_{i+1}b_{i+1}|$

(이차기울기 의 변화) $|(a_{i+1} + a_{i-1} - 2a_i) - (b_{i+1} + b_{i-1} - 2b_i)|$

따라서 3개의 유전자에 대하여 총 12가지의 특성들이 사용 되였다

3.2 은닉격자 모형(SOLL) 을 이용한 분석

은닉격자 모형은 은닉변수의 사전 분포를 입력 데이터의 분포에 따라서 은닉격자를 이용하여 개선시켜 줌으로서 데이터의 분석 및 시각화에 효과적이다. 은닉격자 모형에서 입력 데이터는 은닉격자를 이용하여 은닉공간 상에 표시되고 입력 데이터와 은닉격자를 구성하는 노드 사이의 연결가중치에 따라서 은닉격자의 위치가 정해진다. 은닉격자 모형에 적용된 알고리즘은 EM 알고리즘[1] 이며 표1 과 같다.

E-step 에서는 연결가중치와 은닉노드간의 조건부 확률을 이용하여 은닉노드의 위치가 갱신되고

$$x_k \leftarrow \sum_{i=1}^{K} p(v_k \mid x_i) x_i, \qquad k = 1,..K$$

입력데이터의 사후확률을 이용하여 입력데이터의 은닉공간상의 사상값이 구하여 진다.

$$h_n \leftarrow \sum_{k=1}^{K} x_k p(x_k \mid y_n)$$

[표1] 은닉격자 모형의 계산을 위한 pseudo-code

입력 : 입력데이터 $y_1, y_2, \cdots, y_N$

출력 : 입력데이터에 대한 은닉 변수 값 $h_1, h_2, \cdots, h_N$

초기화 :

- M 개의 RBF 기저함수,
- K 개의 은닉 노 드, $x_1, x_2, \cdots, x_K$
- K 개의 연결가중치, $v_1, v_2, \cdots, v_K$
- 학습 율 η , 페널티 상수 λ , 오차 항 βI , 모델 상수 W

반복 :

(E-step)

- 연결가중치를 이용한 은닉 노 드 x_i 의 갱신

  $x_i \leftarrow \sum_{j=1}^{K} p(v_i \mid x_j) x_j, \quad k = 1,..K$

- 각 입력데이터 y_n에 대한 노 드 x_k의 사후 확률 갱신

  $p(x_k \mid y_n) \leftarrow \frac{p(y_n \mid x_k)}{\sum_{k=1}^{K} p(y_n \mid x_k)}, \quad k = 1,..,K;\ n = 1,..,N$

- 입력데이터 y_n의 은닉 변수 값 h_n의 갱신 $h_n \leftarrow \sum_{k=1}^{K} x_k p(x_k \mid y_n)$

(M-step)

- 모델 상수 W 의 계산 $W \leftarrow [\Phi' RR' \Phi + \lambda I]^{-1} \Phi' RY$
- 각 입력데이터 y_n에 대하여 winner 노 드 x_w 의 인덱스 선택

  $w \leftarrow \arg\min_k \| y_n - v_k \| \quad (n = 1,..,N;\ k = 1,..,K)$

- 연결가중치 v_k 의 갱신

  $v'_k \leftarrow v_k + \eta H(x_k, x_w)(v_k - y_n) \quad (n = 1,..,N)$

$$p(x_k \mid y_n) \leftarrow \frac{p(y_n \mid x_k)}{\sum_{k=1}^{K} p(y_n \mid x_k)}, \qquad k = 1,..,K;\ n = 1,..,N$$

M-step 에서는 앞서 구하여진 은닉변수값을 이용하여 상수 행렬 W를 구한다.

$$W \leftarrow [\Phi' RR' \Phi + \lambda I]^{-1} \Phi' RY$$

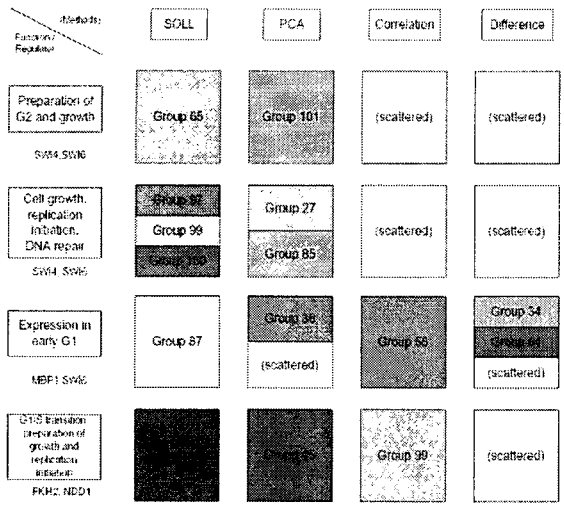
또한 연결가중치 v_k의 값을 입력데이터와 근방함수 H를 이용하여 갱신한다.

$$v'_k \leftarrow v_k + \eta H(x_k, x_w)(v_k - y_n) \qquad (n = 1,..,N)$$

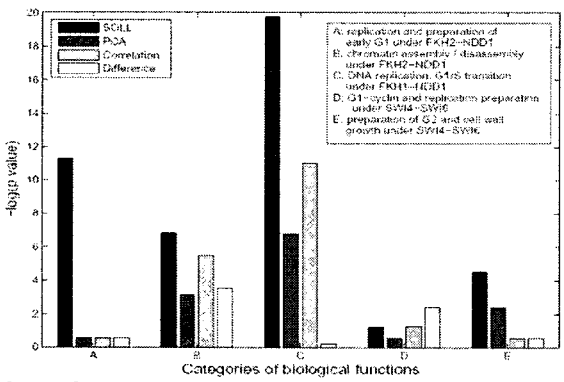
4. 실험 결과

이스트 cell-cycle 과정에 있어서 조절작용에 연관된 생물학적 기능과의 상호 연관성을 찾기위하여 spellman[4] 의 이스트 데이터중에서 53개의 유전자를 선택하고 이중에서 3개의 유전자로 구성된 모든 가능한 유전자군들을 생성하여 실험에 사용하였다. 앞서 정의된 조절작용을 반영하는 값들을 정해서 각 유전자군의 특성을 은닉 공간상에 사상한 결과 유사한 생물학적 기능이나 발현시간의 특성으로 유전자군들을 식별할수 있었다.

은닉격자 모형과의 비교를 위하여 PCA, 상관관계 및 절대값의 차이에 의한 성분들을 이용한 분석결과를 그림1에 표시 하였다. 은닉격자모형 (SOLL)은 조절유전자 SWI4 와 SWI6 와 연관된 유전자군에 대하여 G2기의 전사인자 및 세포 성장에 관여하는 유전자(FKH1, FKH2, NDD1)들을 분류 하였고 (65번 군집). PCA 방법도 비슷한 결과를 보여 주었다. 반면에 상관관계 및 절대값 차이에 의한 분석은 산재된 결과를 보였다. 조절유전자 MBP1, SWI6 와 연관된 유전자군에 대해서 은닉격자모형은 초기 G1기에서의 발현에 관여하는 유전자 (FUS1,HO,



[그림 1] SOLL 및 기타방법에 의한 유사 조절유전자 군의 분류



[그림 2] p-value를 통한 SOLL 및 다른 방법에 대한 성능비교

KAR4)들을 분류 하였고 (87번 군집), PCA 방법과 절대값 차이에 의한 분석은 산재된 결과를 보인 반면 상관관계에 의한 분석은 비슷한 결과를 보였다.

또한 조절유전자 FKH2 와 NDD1에 관한 유전자군에 대해서 은닉격자모형 및 PCA, 상관관계에 의한 방법은 G1/S 기 에서의 전사, 성장과 복제시작과정의 준비에 관여하는 유전자 (CDC45, CDC6, CLB5) 들을 분류한 반면, 절대값 차이에 의한 분석은 산재된 결과를 보였다. 특히 SBF 단백질에 의해 조절되는 유전자군에 대하여 은닉격자모형은 DNA 복제 및 G1기에서의 세포의 초기 성장에 관여하는 유전자를 3개의 군으로 세분하여 분류 하였다. 즉 CDC6, FUS1, PCL9 (군집 99) 등은 세포성장과 DNA 복제준비에 관여하는 초기 G1기에서 발현되는 단백질들을 포함하고 있고 CDC45, CLB5, DPB2 (군집 97)등은 DNA 복제와 극화된 세포의 성장에 필요한 단백질들을 포함하고, RAD27, RFA2 (군집 100) 등은 DNA 재결합및 수리를 위한 단백질들을 포함한다. 다른방법에 의한 결과를 보면, PCA 방법은 두개의 군집으로 분류하였고, 상관관계 및 절대값 차이에 의한 분석은 산재된 결과를 보였다.

그림 2 에서는 은닉격자모형외에 PCA 방법, 주변의 상관관계, 발현값의 차이에 의한 방법에 따른 군집화 결과들을 선택된 생물학적 기능에 따른 p-value 값에 의해 비교하였다. 은닉격자모형에 의한 방법이 " SWI4 와 SWI6 에 관련한 G1 cyclin 및 복제 준비" 기능을 제외한 나머지 기능에서 높은 유의성을 보인 반면 발현값의 차이에 의한 방법이 그 기능에서는 높은 유의성을 보였다.

5. 결론 및 논의

이스트 cell cycle 에 관련된 유전자들로부터 각 시간에 따른 집합적 표현특성을 분석하므로서 조절작용과 관련된 의미있는 유전자 군들을 얻을 수 있었다. SWI4/SWI6 와 MBP1/SWI6 단백질은 G1/S 조절 전사 인자로서 제시된 은닉 격자모형에 의하여 G1/S 에 관여하는 유전자군들이 좀더 세분화되어졌고 G2 기에 해당하는 유전자군들은 좀더 광범위하게 분류 되었다.

은닉변수학습 모형은 고차원의 DNA 발현에 따른 입력값들로부터 은닉변수들을 가정하여 각 유전자군으로부터 특징적인 패턴을 나타내어 준다. 따라서 혼재되어 나타나는 내부적 특징을 재구성하여 보여줄수 있다.

본 논문에서는 세개의 유전자로 이루어진 유전자군들을 대상으로 하였으나, 더 많은 유전자로 이루어진 유전자군의 경우 적절한 생물학적 방법에 의하여 경우의 수를 크게 줄여줄 필요가 있다.

6. 감사의 글

본 논문은 BK21IT, IMT2000 Bioinformatics, NRL program 에 의해 지원 받았음

7. 참고문헌

[1] Dempster, A.P., et. al. " Maximum likelihood from incomplete data via the EM algorithm (with discussion)." *Journal of the Royal Statistical Society*, Series B, **39**, 1-38. (1977)

[2] Dhillon, I.S., et. al. " Diametrical clustering for identifying anti-correlated gene clusters Bioinformatics." *Bioinformatics*, **19**, pp.1612-1619. (2003)

[3] Qian, J., et. al. " Beyond synexpression relationships: Local clustering of time-shifted and inverted gene expression profiles indentifies new, biologically relevant interactions." *J. Mol. Biol.*, **314**, pp. 1053-1066. (2001)

[4] Spellman, P.T., et. al. " Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae. " *Mol. Biol. Cell*, **9**, pp. 3273-3297. (1998)

[5] Zhang B.T., et. al. " Self-organizing latent lattice models for temporal gene expression profiling." *Machine Learning*, **52**, pp. 67-89. (2003)