저차원공간으로의 매핑에 기반한 DNA 서열 요소 및 유전자 발현 패턴간 관련성 분석

이 종 우° 장 병 탁
서울대학교 컴퓨터공학부
{jwlee°, btzhang}@bi.snu.ac.kr

Linking DNA Sequence Motifs with Gene Expression Patterns Based on a Low-Dimensional Mapping

Jongwoo Lee° Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

마이크로 어레이(micro array)로 표현되는 유전자 발현 패턴(gene expression pattern)들과 해당 유전자의 upstream에 위치한 DNA 서열 요소(motif)들은 유전자 발현에 밀접한 관련을 맺고 있는데 이들간의 매핑관계를 알아내는 것은 생물전산학 분야에서 중요한 문제 중 하나이다. 본 고에서는 유전자 발현 패턴 데이터와 해당 DNA에 포함된 것으로 알려진 모티프 프로파일에 대해 대응분석(correspondence analysis)을 수행하고 2차원 평면에 매핑하여 특정 유전자 발현과 밀접하게 관련된다고 여겨지는 후보 모티프를 시각적이고 직관적으로 동정하는 방법을 제시한다. 또한 유전자 발현 패턴은 일정한 길이로 나누어 가능한 모든 패턴에 대해 클러스터링을 행하여 이에 대한 인덱스로 데이터를 표현하여 패턴의 인식성과 발현 순차성을 높이는 반면 복잡도를 줄이도록 하였다. 실험에서 두가지 형태의 모티프 프로파일과 효모 *Saccharomyces cerevisiae* 포자형성 데이터 집합에 대하여 대응 분석을 통한 시각화된 결과를 이용해 유전자 발현과 깊게 관련되는 것으로 알려진 모티프들이 대응 유전자 발현과의 상관성이 잘 동정되고 있음을 알 수가 있다.

1. 서 론

DNA 마이크로어레이 데이터는 다양한 환경 조건하에서 세포나 조직 내의 수천 개 유전자의 발현 양상을 동시에 제공하는데, 이러한 대규모 유전자 발현 데이터 분석에 있어 중요한 이슈 중의 하나는 유전자들의 전사 조절 기작을 이해하는 것이다. 이러한 유전자 발현 조절의 핵심 중 하나는 전사 개시단계에서 전자조절인자(TFs: transcription factor)들이 유전자의 프로모터 영역의 특정 서열요소(motif)와 결합함으로써 이루어지는 것이다. 따라서, 전사조절인자의 결합 사이트(TFBS)로 동작하는 이러한 특정 DNA 서열 요소(이하 모티프)들을 특정 발현 패턴과 연관지어 동정하는 것은 전사 단계에서의 유전자 발현 조절을 이해하는데 있어 중요한 전초 정보를 제공할 수 있다.

이에 대한 잘 알려진 연구중 하나는 발현 패턴이 유사한 유전자들을 클러스터링(clustering)한 후에 동일 클러스터의 유전자들이 upstream에 가지는 모티프들을 찾아내는 방식이다[1, 2, 3]. 이 방식은 발현패턴에 의해서만 클러스터링을 수행하였기 때문에 같은 클러스터내의 유전자들중 모티프를 공유하지 않는 경우와 같은 모티프를 가지고도 발현하지 않는 유전자들이 존재하기도 하는 경우에는 적용되지 못한다. 이런 단점을 극복하기 위하여 Segal은 발현 패턴과 모티프 집합에 대한 클러스터링을 교대로 반복하여 양쪽의 정보를 동시에 이용하는 방법을 제시하였다[4]. 앞의 두가지 방식과는 좀 다르게 모티프 프로파일(profile)로부터 유전자 발현 패턴으로의 회기(regression) 방식은 모티프와 발현 패턴간의 매핑(mapping) 함수를 직접 학습하는 방식으로 선형 회기 모델을 이용하거나[5, 6] 회기트리(regression tree)를 이용하였다[7].

이 논문에서는 유전자 발현과 연관된 모티프를 시각화(visualization)하여 동정(identification)하기 위해 선형분석 방법중 하나인 대응분석(correspondence analysis)을 이용한다. 대응 분석은 분할표 자료의 행과 열 범주를 저차원 공간상의 점들로 동시에 나타내어 그들의 본 연구에서는 발현 데이터를 일정한 크기의 벡터로 재구성시켜서 클러스터링한 후에 이를 모티프 프로파일에 대한 빈도수로 표현하여 대응분석을 수행하였다.

2. 대응분석 기반의 저차원 공간상 데이터 분석

대응분석(correspondence analysis)은 두 개 혹은 n-차원(n-dimensional) 관측벡터(observable vector)의 공기 빈도수(co-occurrence frequency)를 가지는 분할표에 대해서 각 행벡터와 열벡터를 저차원(주로 2차원) 평면상의 한 점에 대응시켜서 상호관계를 탐구하려는 탐색적 다변량 자료분석기법(multivariate data analysis)이다. 대응된 두 점이 서로 가까우면 상관성을 보이며 이는 행벡터-행벡터간 혹은 열벡터-열벡터간의 관계에서도 성립하지만 행벡터-열벡터간에서는 벡터의 방향을 보고 상관성을 추론할 수가 있다. 서로 다른 범주를 표현하는 연관된 2개의 벡터 변수의 집합이 있을 때 하나의 변수(J-차원)를 행벡터들로 구성하고 다른 하나의 변수(I-차원)를 열벡터로 대응시켜 x_{ij}를 성분으로 갖는 행렬을 $\boldsymbol{X}$ (분할표; contingency table)라 하자. 분할표를 이용하여 다음의 행렬들을 도출한다. 여기서 아랫첨자는 행렬의 크기를 표시한다.

$$P_{(I,J)} = \frac{1}{n} X_{(I,J)},$$

$$r_{(I,1)} = P_{(I,J)} 1_{(J,1)},\ c_{(J,1)} = P^{T}_{(I,1)} 1_{(I,1)},$$

$$\widetilde{P} = P - rc^{T},$$

$$D_r = diag(r_1, r_2, \cdots, r_I),\ D_c = diag(c_1, c_2, \cdots, c_J).$$

위의 행렬들을 이용하여 아래의 P^*행렬을 정의한 후 singular value decomposition (SVD)를 수행하면 다음의 U, V 행렬을 얻을 수 있다.

$$P^{*}_{(I,J)} = D_r^{-1/2}{}_{(I,J)} \widetilde{P}_{(I,J)} D_c^{-1/2}{}_{(J,J)}$$
$$= U_{(I,J-1)} \Lambda_{(J-1,J-1)} V^{T}_{(J-1,J)} \quad \cdots\cdots\cdots (1)$$

여기서 대각행렬 Λ는 고유치(eigenvalue)를 대각성분에 가지는데 위쪽 행이 큰 값을 갖도록 정렬한다. 이는 차원축소를 할 때 분산값

(variance)이 큰 축에 대해 상위 행에 해당하는 고유치을 선택하는 것과 동일하게 취급되도록 해준다.

$\tilde{U}=(D)_r^{1/2}U,\ \tilde{V}=(D)_c^{1/2}V$로 잡으면 이는 $\tilde{F}$ 에 대한 SVD 분할 결과가 되며 이로부터 다음 Y, Z 행렬을 구하면 이는 각각 F 의 각 행벡터들과 열벡터들의 분산값이 큰 축에서부터의 새로운 좌표값들을 포함하게 된다.

$$Y_{(I,J-1)}=D_r^{-1}{}_{(I,I)}\tilde{U}_{(I,J-1)}\Lambda_{(J-1,J-1)} \quad \cdots\cdots (2)$$

$$Z_{(J,J-1)}=D_c^{-1}{}_{(J,J)}\tilde{V}_{(J,J-1)}\Lambda_{(J-1,J-1)} \quad \cdots\cdots (3)$$

식 (2)와 식 (3)으로 얻은 Y, Z 두 종류의 벡터집합에 대해 최좌측 2열만 취하여 각 2차원 행벡터들을 하나의 평면에 도시(plotting)하게 되면 2차원 공간으로의 차원축소가 이루어진다. 분할표를 구성하는 행벡터들과 열벡터들의 연관성의 정도를 알아보기 위해 χ^2 통계량(statistic)을 이용할 수가 있는데 χ^2 확률변수는 정규분포를 따르는 확률변수 O_{ij}로부터 식 (4)와 같이 정의되며 식(1)에서 정의된 P^*행렬와 성분과의 관계는 다음과 같다.

$$\chi^2=\sum_{i,j}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}=n\sum_{i,j}p_{ij}^{*2} \quad \cdots\cdots (4)$$

식 (1)과 식 (4)에 의해 χ^2 통계량을 조사해 봄으로써 공기빈도를 나타내는 데이터에 대한 대응분석의 유의성을 알 수 있음을 보여준다.

3. 대응분석에 의한 DNA 서열요소와 발현패턴간 관계 분석

대응분석은 각 행벡터와 열벡터의 순서와는 무관한 선형분석이므로 이 실수값을 이산화(discretization)한 후 분할표의 행벡터에 대한 인덱스로 이용하는 것은 발현 패턴의 순차적 정보를 읽는 결과를 낳는다. 발현 패턴의 순차 정보를 활용하기 위해 여기서는 일정한 크기를 갖는 세그먼트를 슬라이딩 윈도우(sliding window)형식으로 구분하여 각 세그먼트(segment)로 나뉘어지는 벡터를 하나의 스칼라값으로 표현하였다. 이를 위해 전체 발현패턴들에 대해 나올 수 있는 모든 세그먼트들을 클러스터링하여 이에 대한 순차적 인덱스로서 유전자 발현패턴을 재구성하였다[그림 1]. 이 작업은 실수값의 이산화(discretization)과 함께 데이터 압축(compression)의 효과를 동시에 얻게 된다. 이 사전작업에서 결정할 수 있는 인자(parameter)는 세그먼트의 크기, 클러스터링 모델, 그리고 클러스터의 개수가 된다. 여기서 클러스터 개수는 세그먼트 크기와 밀접한 관계가 있을 것으로 추측된다. 대응분석은 발현패턴 프로파일과 모티프 프로파일에 의해 구성된 분할표를 이용하게 된다. 여기서 발현패턴 프로파일은 특정 시각(m개)에 특정 클러스터(n개)에 대한 인덱스(mn개)로 구분하며 클러스터 인덱스는 후보 모티프에 대한 인덱스를 쓰거나 가능한 후보 염기서열에 대한 인덱스로 이용한다. 이때 분할표의 (i, j)성분은 i번째 모티프 프로파일에 해당하면서 동시에 j번째 발현 프로파일을 갖는 유전자들의 총 개수가 된다.

4. 실험 및 결과

실험에 이용될 모티프 프로파일은 2가지 형태를 가지고 이에 따라 2개의 실험모드로 나뉜다. 첫 번째 모드에서 쓰는 모티프 프로파일은 알려진 정제된 모티프 집합에 대한 인덱스로 구성되며 어떤 유전자 발현 패턴이 어떤 모티프와 상관성을 갖는지 동정하는 실험이고 두 번째 모드에서는 유전자 upstream 영역(800bp)에 대해 가능한 모든 연속된 6-mer 염기서열을 구분하는 $4^6(=4096)$가지의 인덱스로 구성된다. 특히 두 번째 실험 모드는 노이즈가 많지만 새로운 모티프의 염기서열을 찾아낼 수 있는 실험 세팅에 해당한다.

4.1 데이터 집합

두가지 실험 모드에 공통으로 이용되는 유전자 발현 데이터는 효묘 *Saccharomyces cerevisiae*의 포자형성(sporulation) 데이터[7]이다. 이 데이터는 meiosis와 포자형성 단계동안 유전자 발현을 6시점(0h, 0.5h, 2h, 5h, 7h, 9h, 11.5h)에 대해 실수화된 것이다. 우리는 6000개의 유전자 발현 데이터 중 변화 수치가 미미한 것과 upstream에 모티프를 가지지 않는 것들 제외한 1865개의 발현 데이터만을 이용하였다. 첫 번째

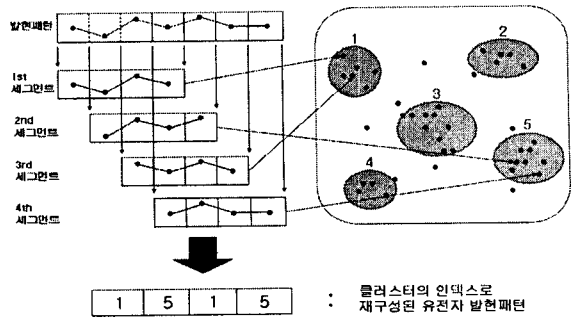

그림 1 유전자 발현패턴의 순차적으로 표현된 클러스터 인덱스로의 재구성.

실험을 위한 모티프 프로파일의 인덱스로 쓰인 모티프 집합은 Pilpel에 의해 정리된 효모 *Saccharomyces cerevisiae* 유전자 발현과 관련되어 있는 356개로 이루어졌다. 이들중 37개는 유전자 발현과 밀접하게 관련되어 있다고 알려져 있다.

4.2 연관성의 유의성 분석

만들어진 분할표로부터 DNA 서열 요소와 발현 패턴간의 연관성의 유의성을 분석해 보았다. χ^2 검증 결과, p-value는 1.55×10^{-98}이었는데 이는 그 연관성의 정도가 통계적으로 상당히 유의함(significant)을 의미한다. χ^2 검증과 더불어 추가적으로 정보이론에서의 엔트로피(entropy)에 기반한 식 (5)의 척도에 의해 모티프 및 유전자 발현 패턴의 연관성의 정도를 측정해 보았다. ($I(x,y)$: mutual information, $H(x)$: entropy)

$$U(x,y)=2\left[\frac{I(x,y)}{H(x)+H(y)}\right] \quad \cdots\cdots (5)$$

주어진 데이터에 대해 $U(x,y)$는 각각 0.027이었으며, 상대적으로 비교해 볼 때 1000번의 임의 매칭(random matching)에 의한 분할표 구성들 중에서 원래의 데이터에 대한 경우보다 높은 값을 갖는 경우는 한 번도 나타나지 않았다. 결론적으로, 두 척도에 의한 검증 결과 모티프 프로파일과 유전자 발현 패턴간의 연관관계는 상당히 의미있는 수준이라고 할 수 있다.

4.3 유효한 모티프 파악 및 유전자 발현패턴과의 관련성 시각화

실험을 결정하는 인자로서 우리는 유전자 발현패턴의 세그먼트 크기를 4로 클러스터 개수를 20으로 모티프의 염기서열 크기를 6-mer로 정하였으며 발현 패턴의 순차적 클러스터 인덱스 재구성을 위해 k-means 클러스터링 모델을 사용하였다.

표 1은 첫 번째 실험 모드의 결과로서 대응 분석에 의해 저차원 좌표로 매핑된 좌표 중 각 축에 대해 절대값이 제일 높은 순으로 정렬된 최대 10개의 모티프들의 목록이다. 대응 분석은 χ^2통계량과 밀접한 관련이 있으며, 따라서 이 모티프들은 유전자 발현 패턴과의 연관성이 큰 모티프들이라고 할 수 있다. 그림 2는 대응 분석 결과로부터 도출된 두 축에 의해 정의된 2차원 공간상에 이 모티프들과 발현 패턴들을 도시한다. 표시된 발현 패턴들은 모티프와 마찬가지로 각 축에 대해 가중치의 절대값이 큰 상위 10개에 해당한다. .MSE, URS1, MCB, PAC, RAP1 등은 기존의 연구에서 잘 알려진 모티프들이다. URS1은 Ume6가 결합하는 중요한 모티프로서, URS1은 효모의 포자형성 초기 단계에 관여하는 유전자들의 upstream 영역에서 자주 관찰된다[8, 10]. 실험에 사용된 마이크로어레이 데이터에서 두 번째(0.5h)와 세 번째(2h) 샘플이 감수분열 전기에 해당된다. MSE는 전사인자 ndt80이 결합하는 모티프로 알려져 있으며, ndt80은 감수분열(meiosis) 전기(prophase)의 종료 시점에 middle sporulation 유전자들의 발현을 증진시키는 전사인자로 알려져 있다 [8, 11]. 그림 2에서 MSE와 URS의 좌표상의 위치 및 이들과 가깝거나 비슷한 방향의 유전자 발현 패턴들을 살펴보면, 대응분석에 의한 결과가 이러한 사실을 적절히 반영하고 있으며 이들 모티프를 upstream 영역에 갖는 두 유전자 집합들의 전사

1번째 축	PAC, mRRPE, m_RRSE3, m310, SFF', m281, MCB, ndt80(MSE), Ume6(URS1), m190
2번째 축	ndt80(MSE), SFF, SFF', m190, Ume6(URS1), m310, m_RPE72, m_RPE69, m_RPE6, RAP1

표 1 최대 고유값을 갖는 축에 대한 최대 좌표값을 갖는 모티프와 유전자 발현 패턴 리스트.

적 조절 양상 차이가 두 번째 샘플(05.h)과 세 번째 샘플(2h)에서의 발현 패턴에 있음을 알 수 있다. 모티프 m190은 MIPS의 meiosis 기능목록(functional category)에 속한 유전자들의 집합에서 추출된 서열로서, 서열 패턴면에서 URS1과 상당히 유사하며 두 모티프를 갖는 유전자들의 집합은 아주 유사한 발현 패턴을 보였다. MCB는 [8]의 연구 결과에서는 명시적으로 제시되지 않았지만, [5, 7]의 연구에서는 유의미한 모티프로 제시되었다. PAC, mRRPE에 관련된 유전자 발현 패턴은 전체적으로 억제된 발현패턴을 보였으며, 특히 [9]에서는 이 두 모티프가 유전자 발현 조절에 있어 서로 상승작용을 일으키는 것으로 제시되었다. RAP1 역시 전체적으로 억제되는 발현패턴과 연관있는 것으로 제시되었지만, 0.5h(두번째 샘플) 지점에서의 발현 억제 후 그 회복이 좀 더 지연된다는 점에서 조금 차이가 있다. PCA, mRRPE, RAP1과 더불어 SFF는 [7]에서도 중요한 모티프로 동정되었다. 이와 같이, 모티프와 발현 패턴간의 공기 정보의 대응 분석을 통한 저차원공간상에서의 분석을 통해 서로 연관 정도가 큰 모티프 및 발현 패턴을 효과적으로 추출할 수 있으며, 또한 이의 가시화를 통해 둘 간의 관계를 보다 직관적으로 제시할 수 있음을 알 수 있다. 염기서열을 바탕으로한 raw sequence 모티프 프로파일에 대한 실험 결과는 그림 3에 제시되어 있다. 이 경우에도 그림 2에서와 같이 비슷한 결과를 보임을 알 수 있으며, 따라서 논문에서 제시된 방법이 후보 서열 모티프 선택에서 발생할 수 있는 노이즈에 비교적 강건하다고 할 수 있다.

5. 결론

본 논문에서는 DNA 서열요소와 유전자 발현 패턴의 공기 정보의 저차원 공간상 분석에 의한 특정 발현 패턴에 대응되는 의미있는 DNA 서열 요소의 동정을 제시하였다. 실험 결과에서 보듯이 모티프와 유전

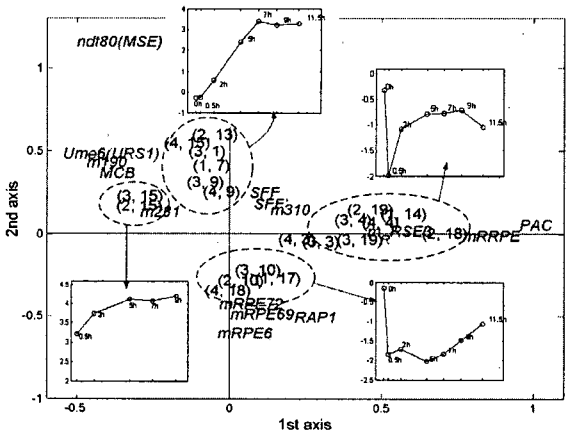

그림 2 대응 분석을 통해 단일 2차원 평면에 도시된 모티프와 유전자 발현 패턴.

자 발현 패턴과의 공기 빈도수에 의해 만들어진 분할표 데이터는 χ^2 통계 검사 및 엔트로피를 이용한 상관성 측정을 통해 높은 유의성을 가지는 것을 알 수 있었으며 대응분석을 수행하여 2차원 공간에 매핑

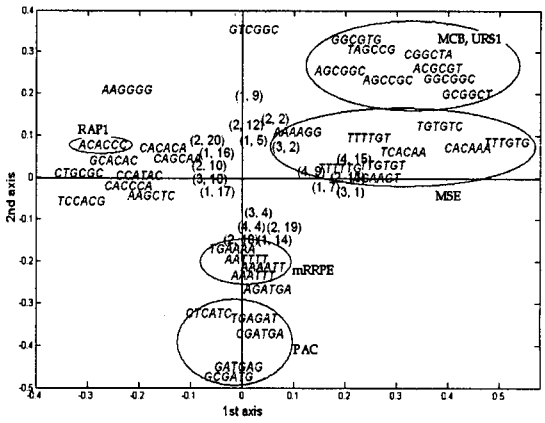

그림 3 유전자 upstream 영역의 6-mer 염기서열 데이터에 대한 대응분석 결과.

시킨 이들의 프로파일을 통해 이들의 상관성을 쉽게 시각적으로 쉽게 인지할 수가 있었다. 차후에 실험 결정 요인들의 변화에 대한 최적 분할표 구성과 저차원 공간으로의 매핑에 관한 다른 기법들과의 비교 연구들이 필요할 것으로 보인다.

6. 감사의 글

이 논문은 교육부 BK21사업과 과학기술부 국가지정연구실사업(NRL) 및 산업자원부 차세대 신기술과제에 의하여 지원되었음.

7. 참고 문헌

[1] P. T. Spellman et al., "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization", *Molecular Biology of the Cell*, 9, 3273-3279, 1998.

[2] F. Roth, P. Hughes, J. D. Estep and G. Church, "Finding DNA regulatory motifs within unaligned noncoding sequenesclustered by whole-genome mrna quantification", *Nature Biotechnology*, 16, pp. 939-945, 1998.

[3] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture", *Nature Genetics*, 22, pp. 281-285, 1999.

[4] E. Segal, R. Yelensky, and D. Koller, "Genome-wide discovery of transcriptional modules from DNA sequence and gene expression", *Bioinformatics*, 19(Suppl. 1), pp. i273-i282, 2003

[5] H. Bussemaker, H. Li, and E. D. Siggia, "Regulatory element detection using correlation with expression", *Nature Genetics*, 27, pp. 167-171, 2001.

[6] S. Keles, M. J. van der Laan, and M. B. Eisen, "Identification of regulatory elements using a feature selection method", *Bioinformatics*, 18(9), pp. 1167-1175, 2002.

[7] T. M. Phuong, D. Lee, and K. H. Lee, "Regression trees for regulatory element identification", *Bioinformatics*, 20(5), pp. 750-757, 2004.

[8] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, and I. Herskowitz, "The transcriptional program of sporulation in budding yeast", *Science*, **282**, pp. 699-705, 1998.

[9] Y. Pilpel, Sudarsanam, P., and G. M. Church, "Identifying regulatory networks by combinatorial analysis of promoter elements", *Nature Genetics*, 29, pp. 153-159, 2001.

[10] R. M. Williams, et al., "The Ume6 regulon coordinates metabolic and meiotic gene expression in yeast", *PNAS*, 99(21), pp. 13431-13436, 2002.

[11] M. Pierce, et al., "Sum1 and Ndt80 proteins compete for binding to middle sporulation element sequences that control meiotic gene expression", *Molecular and Celluar Biology*, 23(14), pp. 4814-4825, 2003.