Kernel CCA를 이용한 유전자 발현 조절 기능 모티프 추출

이제근[O1,2] 정제균[1,2] 장정호[3] 장병탁[1,2,3]
서울대학교 생물정보학 협동과정[1]
서울대학교 바이오정보기술 연구센터[2]
서울대학교 컴퓨터공학부[3]
{jkrhee[O], jgjoung, jhchang, btzhang}@bi.snu.ac.kr

Identification for Gene Regulatory Motifs by Kernel CCA

Je-Keun Rhee[01,2] Je-Gun Joung[1,2] Jeong-Ho Chang[3] Byoung-Tak Zhang[1,2,3]
Graduate Program in Bioinformatics, Seoul National University[1]
Center for Bioinformation Technology, Seoul National University[2]
School of Computer Science and Engineering, Seoul National University[3]

요 약

유전자 발현은 많은 전자조절인자에 의해서 조절된다. 이러한 조절인자들은 각각 DNA 상에 존재하는 특정한 모티프에 결합하여 그 기능을 수행한다. 따라서 DNA 상의 특정한 서열 정보가 유전자 발현과 직접적으로 연관되어 있다고 생각할 수 있다. 본 논문에서는 두 가지 서로 다른 데이터들에 대한 관계를 알아보기 위하여 사용되는 방법인 Kernel CCA를 이용하여 DNA 상의 특정한 모티프와 유전자 발현 사이의 관계를 알아보았다. 이를 이용한 실험 결과, 유전자 발현과 밀접하게 관련되어 있는 모티프들을 발견할 수 있었고, 기존에 중요한 것으로 알려져 있는 모티프가 실제로도 유전자 발현에 밀접한 영향을 미친다는 것을 알 수 있었다.

1. 서 론

생물학 연구는 생물학적, 혹은 화학적인 실험을 통해서만 가능한 것으로 생각되어 왔다. 하지만 최근 들어 마이크로어레이(microarray) 등의 기술이 발전되면서 수많은 데이터를 동시에 분석할 필요성이 생기게 되었고, 이로 인해 생물학 연구에서도 컴퓨터를 이용한 분석이 필수적이 되었다. 유전자 발현 및 조절 기작과 같이 오래전부터 생물학 연구실에서 중점적으로 연구해오던 주제에서도 이와 같은 연구 방법의 변화는 나타나고 있다. Spellman 등은 마이크로어레이를 통해 얻은 6000여개의 모든 효모(yeast) 유전자의 발현 양상을 이용하여 세포주기(cell cycle)에 관련되어 있는 유전자들에 대한 분석을 수행하였다[1]. 또한 Spellman의 데이터(data)를 이용하여 SOM (self-organized map)을 이용하여 유전자들의 군집화(clustering)를 통한 유전자 발현과 관련된 연구를 수행한 결과도 존재하는 등[2] 유전자 발현 및 조절 기작 등과 관련하여 계산학적(computational) 방법을 통한 많은 연구가 수행되고 있다.

하지만 많은 연구에도 불구하고 유전자 발현과 연관되어 있는 것으로 추정되는 유전자 상류 영역(upstream region) 서열과 실제 유전자 발현 양상과의 직접적인 관계에 대해서는 아직 명확하게 밝혀지지 않고 있다.

유전자 발현은 RNA 중합효소(RNA polymerase)와 함께 실질적인 전사(transcription) 과정을 조절하는 많은 전사 인자(transcription factor, TF)들이 상류 영역(upstream region)에 결합하면서 일어나게 된다. 전사인자들이 상류 영역의 TFBS (transcription factor binding site)라고 불리는 영역에 결합함으로서 유전자 발현이 조절되는 것이다. DNA 상에서 전사인자들이 결합할 수 있는 특정한 서열은 모티프(motif)라고도 불린다. 이러한 생물학적인 사실들로부터 모티프와 유전자 발현 사이에 밀접한 연관성이 있다는 것을 추정해볼 수 있다.

본 논문에서는 Kernel CCA (Kernel Canonical correlation analysis)를 이용하여 모티프와 유전자 발현 양상과의 상관관계를 분석해보았다. Kernel CCA는 서로 다른 두 데이터에 대한 관계를 알아보기 위해 사용될 수 있는 방법이다. Kernel을 이용하여 두 데이터를 보다 높은 차원(dimension)의 자질 공간(feature space)으로 사상(mapping)시키고 이때의 각 투영(projection) 성분(component)들 사이의 관계를 측정하는 것이다.

본 논문에서는 우선 모티프와 유전자 발현 사이의 연관 관계를 측정하기 위해 사용된 Kernel CCA 방법에 대해 설명하고, 실제 생물학 데이터에 어떻게 적용하여 분석하였는지를 설명하고자한다. 그리고 그 결과는 유전자 발현과 모티프 사이에 밀접한 관계가 있음을 보여준다. 또한 이때 특히 유전자 발현에 중요한 영향을 미치는 모티프를 찾을 수 있었고, 이 결과는 기존의 논문들에서 알려진 결과들과도 부합된다는 것을 확인할 수 있었다.

2. 실험 방법

2-1. 유전자 발현에 영향을 미치는 모티프를 찾기 위한 Kernel CCA 방법

Kernel CCA는 고전적인 통계학적 분석 방법인 CCA에

kernel 방법을 적용시킨 것이다[3, 4]. CCA는 서로 다른 두 데이터의 최대의 상호 관계(maximal correlation)를 가지는 각 변수들의 선형 결합(linear combination)을 찾는 방법이라고 할 수 있다. 하지만 여기에 kernel 방법을 적용하면 CCA가 가지는 선형성에 대한 한계를 극복할 수 있고, 비선형 데이터에 대한 분석이 가능해질 수 있다. 즉 Kernel CCA의 이러한 특성을 이용하여 유전자 발현 양상과 모티프에 대한 연관 관계를 분석해보고자 하는 것이다.

유전자 발현 데이터와 모티프 데이터는 각각 $x_{exp}=\{e_1, e_2, \ldots, e_N\}$, $x_{motif}=\{m_1, m_2, \ldots, m_M\}$으로 표현될 수 있다. 여기서 e는 각 시간에 따른 유전자 발현양을 의미하며, m은 특정한 모티프를 의미한다. 두 데이터는 ϕ에 의해 Hilbert 공간 H로 사상된다. 이 때 Hilbert 공간상에서의 각 데이터에 대한 방향 f_{exp}, f_{motif} 사이의 최대 상관값(maximal correlation)을 구하는 것이다. f 값은 Hilbert 공간으로의 투영 시켜주는 u로서 표현될 수 있다.

$$u_{\exp} = < f_{\exp}, \phi_{\exp}(x_{\exp}) > \quad (1)$$

$$u_{motif} = < f_{motif}, \phi_{motif}(x_{motif}) > \quad (2)$$

이 식은 Lagrangean을 이용하여 다음과 같은 식을 계산함으로서 그 답을 구할 수 있다.

$$L_0 = E[(u_{\exp} - E[u_{\exp}])E[(u_{motif} - E[u_{motif}]) - \frac{\rho_{\exp}}{2}E[(u_{\exp} - E[u_{\exp}]^2) - \frac{\rho_{motif}}{2}E[(u_{motif} - E[u_{motif}]^2) \quad (3)$$

여기서 Kernel 행렬(matrix)을 이용하면 식 (3)은 다음과 같이 나타날 수 있다.

$$L = \alpha_{\exp}^T K_{\exp} K_{motif} \alpha_{motif} - \frac{\rho_{\exp}}{2}\alpha_{\exp}^T (K_{\exp} + \lambda_{\exp} I)^2 \alpha_{\exp} - \frac{\rho_{motif}}{2}\alpha_{motif}^T (K_{motif} + \lambda_{motif} I)^2 \alpha_{motif} \quad (4)$$

식 (4)에서 I는 단위행렬을 나타낸다. 식 (4)에 대해서 정형화(regularization) 파라미터(parameter) λ를 조정하면서 Lagrangean을 최대로 만드는 것은 다음 식 (5)의 고유값(eigenvalue) 문제를 푸는 것으로 변환될 수 있다.

$$\begin{pmatrix} 0 & K_{\exp}K_{motif} \\ K_{motif}K_{\exp} & 0 \end{pmatrix}\begin{pmatrix} \alpha_{\exp} \\ \alpha_{motif} \end{pmatrix} = \rho\begin{pmatrix} (K_{\exp} + \lambda_{\exp} I)^2 & 0 \\ 0 & (K_{\exp} + \lambda_{\exp} I)^2 \end{pmatrix}\begin{pmatrix} \alpha_{\exp} \\ \alpha_{motif} \end{pmatrix} \quad (5)$$

여기서 각각의 표준 상관 점수(canonical correlation score, CC score)는 $u_{exp}=K_{exp}\alpha_{exp}$와 $u_{motif}=K_{motif}\alpha_{motif}$와 같이 계산되며, 이때 각 자질(feature)들에 대한 가중치(weight) 벡터 W는 다음과 같이 계산될 수 있다.

$$W_{\exp} = X_{\exp}^T \alpha_{\exp} \quad (6)$$

$$W_{motif} = X_{motif}^T \alpha_{motif} \quad (7)$$

만일 데이터 X에서 특정 속성(attribute)에 대한 가중치가 큰 값을 가진다면, 이 속성은 실제로 유전자 발현과 강한 상호 관련성을 가지고 있는 것으로 해석할 수 있다. 즉 가중치의 절대값이 높은 모티프는 유전자 발현에 강한 영향을 미치는 모티프라고 생각할 수 있는 것이다.

2-2. 실험 데이터 및 설정

본 논문에서는 Pilpel이 추출한 효모의 모티프 정보를 이용하였다[5]. 이 데이터를 바탕으로 AlignACE를 이용하여 효모의 각 유전자가 총 42개의 모티프 중 어떤 모티프들을 가지고 있는지를 얻을 수 있다.

또한 유전자 발현 정보에 대해서는 Spellman에 의한 마이크로어레이 결과를 이용하였다[1]. 이 데이터는 효모의 세포 주기 상에서 각 유전자들의 발현 패턴에 대한 정보를 보여준다. 본 논문에서는 Spellman의 유전자 발현 데이터 중 alpha factor synchronization의 경우에 나온 데이터만을 이용하였다. 이 데이터는 시간의 흐름에 따라 총 18 시점에서 각 유전자 발현양을 측정한 것이다.

한편 본 실험에서는 총 6000여개의 효모 유전자들 중 Spellman의 분석 결과에서 세포 주기에 관련되어 있다고 알려진 것만을 이용했다. Spellman의 실험 결과에 따라 약 800개의 유전자들이 세포 주기에 관련된 것으로 알려져 있다. 이 800개의 ORF에 대한 정보만을 이용하여서도 분석을 수행해볼 수 있다. 하지만 이 중 유전자 발현 값이 없는 데이터들이 일부 존재하여, 이를 제외하고, 총 551개의 데이터를 실제 실험에 사용하였다.

모티프 데이터의 Hilbert 공간으로의 변환에서는 polynomial kernel을 이용하였다.

$$k(x_{\exp}^1, x_{\exp}^2) = (x_{\exp}^1 \bullet x_{\exp}^2 + 1)^d \quad (9)$$

이 때 d(degree)의 값은 3으로 하였다. 유전자 발현 데이터에 대해서는 식 (10)의 gaussian RBF kernel을 이용하였다.

$$k(x_{\exp}^1, x_{\exp}^2) = \exp(-\frac{d(x_{\exp}^1, x_{\exp}^2)}{2\sigma^2}) \quad (10)$$

파라미터 σ의 값은 0.5이며, d는 두 데이터간의 거리를 의미한다. 이 때 Kernel CCA를 이용하여 모티프와 유전자 발현의 두 데이터 간의 관계를 분석하였다.

3. 실험 결과

두 데이터에 대해서 첫 번째 표준 점수(canonical score)에 대한 그래프는 그림 1과 같이 나온다. 각 점은 하나의 유전자에 대한 정보를 나타내며, 각 데이터들이 대각선 방향으로 모여 있다는 것은 모든 유전자들에서 모티프와 유전자 발현 정보가 강한 상관관계를 가지고 있다는 것을 보여주고 있는 것이다.

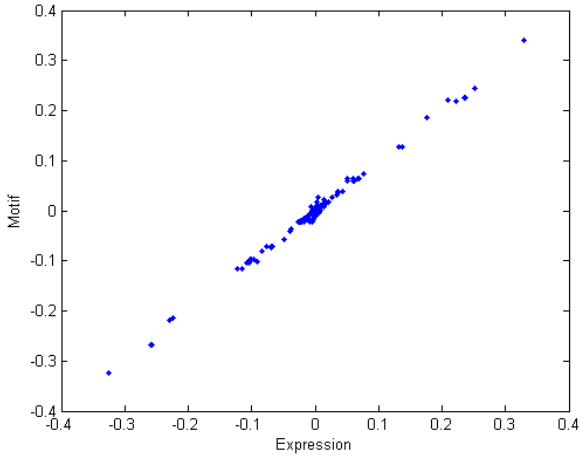

그림 1. CC score를 이용한 유전자 발현 양상과 모티프 정보 사이의 관계 그래프.

식 (7)을 이용하여 각 모티프에 대한 가중치를 계산할 수 있다. 표 1은 높은 가중치 값을 가지는 모티프에 대해 보여주고 있다.

Motif	Weight Score	Function
SWI5	0.89026	G1기에서 발현되는 유전자들의 전사를 활성화시키는 전사 인자 Swi5의 결합 부위
SFF'	0.45399	세포주기 상에서 작용하는 전사 인자 FKH1에 대한 결합 부위
MCB	0.29633	G1기에서 중요하게 작용하는 전사 인자 MBF가 결합할 수 있는 부위
LYS14	0.21796	라이신(lysine) 합성에 관련된 유전자를 활성화시킬 수 있는 단백질이 결합할 수 있는 영역
ALPHA2	0.16532	반수체 세포에서 특정 유전자를 억제하는 단백질에 대한 모티프

표 1. 가중치가 높게 나온 모티프와 그 결과값. 가장 높은 값이 나온 모티프 5개에 대해서만 표에 보여주었다.

표 1에서 보는 것과 같이, 기존의 연구에서 효모의 세포주기에 관련되어 있다고 알려져 있는 모티프들을 Kernel CCA를 통해서 찾을 수 있었다. 대표적으로 MCB와 같은 모티프는 세포 주기 상에서 G1기에서 특히 중요한 역할을 수행하는 것으로 알려져 있다[6]. MCB 모티프에 MBF라는 단백질이 결합하여 이 모티프를 가지고 있는 유전자들의 발현을 조절할 수 있는 것이다. 또한 SWI5 단백질 역시 세포 주기 상에서 중요한 역할을 수행하는 것으로 생각되는 것이며, SFF' 모티프 역시 중요한 조절 단백질인 FKH1이 DNA에 결합하는 영역이다. ALPHA2 모티프의 경우도 기존의 연구 결과에서 세포주기와 관련된 것으로서 생각되는 모티프이다.

이와 같은 방식으로 각 유전자들에 대한 모티프 정보와 유전자 발현 데이터를 Kernel CCA를 이용하여 분석하면 실제 사실과 부합되는 결과가 나온다는 사실을 알 수 있다. 또한 현재로서는 명확히 밝혀지지 않았지만, Kernel CCA에서 유전자 발현 패턴과의 상호 연관성을 계산하는 데에 있어서 높은 점수를 가지는 모티프는, 실제로 유전자 발현에 직접적으로 영향을 줄 수 있는 모티프인 것으로 추정해볼 수 있다.

4. 결 론

본 논문에서는 Kernel CCA를 이용하여 효모 세포주기 상의 유전자 발현 데이터와 모티프 간의 연관 관계를 분석해보고, 이를 기반으로 유전자 발현에 중요한 영향을 미치는 모티프가 무엇인지를 알아보았다. 이 때 나온 결과는 실제 생물학적인 사실과도 부합되므로 매우 의미있는 결과라고 할 수 있겠다.

이번 연구는 복잡한 유전자 조절 과정에서 서열 정보만을 기반으로 하여서도 유전자 발현 양상을 예측해볼 수 있는 기반을 마련했다고 할 수 있겠다. 또한 반대로 유전자 발현 양상만을 알고 있을 때, 이와 같은 패턴이 나타나도록 하는데에 중요한 모티프를 예측해 볼 수 있을 것이다. 이와 같은 예측은 실제로 특정 유전자에 대한 전사 조절 인자를 발견하는 과정으로도 사용될 수 있을 것이다.

현재의 연구는 가중치 벡터를 이용하여 특정 모티프 하나씩만을 기반으로 하여 모티프와 유전자 발현 패턴 사이의 관계에 대하여 분석해보았다. 하지만 Kernel에 대한 기본적인 이해를 바탕으로 하여, 이번 방법에서 사용된 모델을 보다 발전시킨다면 특정 모티프들 간의 조합에 의한 유전자 발현에 미치는 영향도 발견할 수 있을 것이다.

감사의 글

이 논문은 과학기술부 국가지정연구실 사업(NRL)에 의하여 지원되었음.

참고문헌

[1] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell*, Vol. 9, pp. 3273-3297, 1998.

[2] J. Kasturi and R. Acharya, Clustering of diverse genomic data using information fusion, *Bioinformatics*, Vol 21, pp. 423-429, 2005.

[3] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, Canonical correlation analysis; An overview with application to learning methods. Technical Report CSD-TR-03-02, Royal Holloway University of London, 2003.

[4] Y. Yamanishi, J.-P. Vert, A. Nakaya, and M. Kanehisa, Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, Vol. 19 Suppl. 1, pp. i323-i330, 2003.

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

[6] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis, A genome-wide transcriptional analysis of the mitotic cell cycle, *Molecular Cell*, Vol. 2, pp. 65-73, 1998.