인공신경망을 이용한 세포 주기상의 전사 조절 모티프 탐색

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

Transcriptional Regulatory Motif Identification in Cell Cycle using Artificial Neural Networks

Je-Keun Rhee[O1,2] Je-Gun Joung[1,2] Byoung-Tak Zhang[1,2,3]
Interdisciplinary 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

생체 내의 모든 기능은 유전자 발현에 의해 결정된다. 유전자 발현은 많은 인자들에 의해 조절되며, 이러한 조절 과정에 따라 유전자 발현량이 결정되는 것이다. 세포 주기 역시 유전자 발현과 밀접한 연관성을 가지고 있다. 본 논문에서는 효모에서 세포 주기의 각 단계와 관련된 유전자들의 분석을 통해서 세포 주기를 조절하는데 있어서 중요한 역할을 수행하는 전사 조절 모티프들이 무엇인지를 찾아보았다. 주요 모티프의 추출은 인공신경망 모델을 학습하고, 입출력 에러 분석을 통하여 이루어진다. 그 결과 MCB 등 기존의 실험 결과를 통하여 세포주기에 관련이 있다고 알려진 모티프들이 높은 점수를 보인다는 것을 알 수 있었고, 그 외에 세포주기의 각 단계에서 유전자 발현에 중요한 역할을 수행할 것으로 예상되는 다른 모티프들도 예측해볼 수 있었다.

1. 서 론

생체 내에서 세포는 주기적으로 분열하는데, 이를 세포주기(cell cycle)라고 부른다. 세포 주기는 실제로 체세포분열이 일어나는 M기, 단백질의 합성을 준비하는 G1기, 유전자 복제가 일어나는 S기, 세포 분열을 준비하는 G2기로 나누어진다.

이러한 세포 주기의 각 단계에서는 발현되는 유전자들 역시 각각 다르다. 기존의 실험에서 이미 세포 주기 상의 각 단계에 특별히 많은 발현량을 보이는 유전자가 무엇인지를 밝혀내는 실험이 있어왔다[1, 2]. 즉 세포 주기 역시 특정한 유전자의 발현 과정과 직접적인 연관성을 가지고 있는 것이다. 이러한 유전자 발현은 많은 전사 인자(transcription factor)들에 의해 조절된다. 전사 인자들이 DNA상의 특정한 부위와 결합하여 유전자 발현을 조절하게 되는데, 이 전사인자들이 결합할 수 있는 DNA상의 위치를 TFBS(transcription factor binding site) 또는 모티프(motif)라고 한다.

이에 유전자 발현 조절에 중요하게 작용하는 것으로 생각되는 전사 조절 모티프(motif)를 찾아보려는 시도들이 많이 진행되고 있다. 최근에는 회귀나무(regression trees) 및 의사결정나무(decision trees) 모델 등을 이용하여 유전자 발현에 영향을 미치는 모티프를 밝히는 실험도 수행되었다[3, 4]. 하지만 기존의 연구에서는 모델의 크기가 크고 해석이 어렵다는 단점이 있다.

본 논문에서는 단순한 세포 주기 전체가 아닌, 보다 세부적인 세포 주기상의 각 단계별로 영향을 미치는 유전자가 다르다는 사실에 착안하여, 세포 주기상의 각 단계와 관련된 모티프를 밝혀보고자 하였다. 이를 위하여 대표적인 기계학습 방법 중 하나인 인공신경망(artificial neural network, ANN) 모델을 이용하여 실험을 수행하였다. 인공신경망은 기존의 알려진 많은 자료의 학습을 통해 새로운 자료에 대해 적절한 분류(classification)를 가능하도록 해주는 방법으로 알려져있다. 이러한 인공신경망 모델을 이용하여 세포 주기상의 다른 단계에서와는 달리 특이하게 중요한 역할을 수행하는 것으로 생각되는 모티프가 어떤 것인가를 예측해 보았다.

2. 인공신경망을 통한 모티프 탐색 방법

2.1. 인공신경망의 이용

인공신경망은 일반적으로 분류(classification) 또는 회귀(regression) 문제에 적용되어 왔는데 주로 목표 데이터(target data)의 예측(prediction)을 위하여 활용되어왔다. 예를 들어, 입력 데이터 $D = \{x_1, x_2, \cdots\cdots, x_N\}$이 주어지면 목표 데이터 y의 예측을 위하여 신경망 모델이 학습된다. 하지만 여기에서는 목표 데이터에 대한 예측이 아닌 자질 선택(feature selection) 관점에서 중요한 자질(feature)의 추출을 목적으로 신경망을 학습한다. 즉 목표 데이터 y의 예측에 중요하게 작용하는 자질 x_i를 선

정하고자 하는 것이다. 그림 1은 이와같은 목적으로 본 실험에서 사용된 인공신경망에 대한 개략적인 모델을 보여준다.

본 논문에서는 효모(*Saccharomyces cerevisiae*)의 데이터를 가지고 실험하였다. 효모의 각 ORF(open reading frame)에 대한 모티프 정보들이 인공신경망의 입력으로 들어가게 된다. 학습의 결과로는 각 ORF들이 세포주기의 특정 단계에 속하는지 여부가 판단될 수 있다. 이 결과를 통해, 학습을 통한 분류의 정확도가 계산될 수 있다.

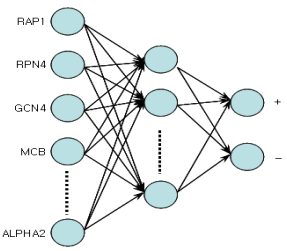

그림 1 ORF의 각 모티프 정보를 입력으로 받는 인공신경망의 기본 모델

다음으로는 전체 *N*개의 입력 데이터에서 특정한 하나의 모티프 x_i가 제외된 *N-1*개의 데이터가 인공신경망의 입력으로 들어가게된다. 따라서 인공신경망 모델의 입력 노드(input node) 역시 *N-1*개가 된다. 이 변형된 모델을 학습하여 나온 분류 정확도를 *N*개의 전체 입력 데이터를 이용하여 만들어진 원래의 모델로부터 나온 결과와 비교하면, 특정 모티프 x_i의 중요도를 확인할 수 있다.

2.2. 모티프 중요도에 대한 점수 계산

본 논문에서 각 모티프에 대한 중요도의 결과값은 다음에 의한 식으로 정의된다.

$$S(M_{motif_i}) = E(M_{total-motif_i}) - E(M_{total}) \quad (1)$$

여기서 M_{total}은 전체 모티프 집합을 의미하며, M_{motif_i}는 i번째 특정 모티프를 의미한다. 따라서 $E(M_{total})$은 전체 모티프 정보를 모두 이용하여 학습하였을 때의 에러율이 되며, $E(M_{total-motif_i})$는 전체 모티프에서 특정한 하나의 모티프 $motif_i$를 제외한 나머지 모티프 정보만을 입력으로 하여 학습하였을 경우의 에러율이 되는 것이다. 마지막으로 $S(M_{motif_i})$는 특정 모티프 $motif_i$에 대하여 본 실험에서 구하고자하는 최종 결과값이 된다.

결국 세포 주기의 각 단계마다, 하나의 모티프 $motif_i$에 대한 정보를 제외하고 학습을 수행하여 나온 결과에 대해 식 (1)을 통하여 최종 점수값을 구하게 된다. 이 때 $S(M_{motif})$가 큰 값을 가질수록, 그 $motif_i$가 세포 주기상의 특정 단계에서 더 중요한 역할을 수행하는 모티프라고 생각할 수 있다.

3. 실험 및 결과

3.1 실험 설정

데이터는 먼저 모티프에 대하여 AlignACE 프로그램을 이용하여 모티프 정보를 추출하여 얻은 Pilpel의 데이터를 사용하였다[5]. 이를 통해 효모의 각 ORF에 대하여 총 42가지의 모티프 포함 여부를 결정할 수 있었다. 세포 주기에 대해 특정 ORF들이 어떤 클래스(class)에 속하는지의 여부는 Spellman의 유전자 발현 패턴 분석을 통하여 얻어진 결과를 토대로 하였다[1]. 이러한 ORF들의 각 단계에서의 모티프 함유 정보가 인공신경망의 입력으로 들어가게 되는 것이며, 이 ORF가 세포 주기상의 특정 클래스에 속하는지 여부에 대하여 학습이 수행된다.

실험에 사용된 세포 주기상의 각 단계별 데이터의 수는 표 1과 같다. 각 단계별로 Spellman에 의해 세포 주기의 특정 단계에서 많은 발현량을 보인다고 밝혀진 ORF에 대한 모티프 데이터들이 모두 positive data로 사용된다[1]. 한편 negative data는 6000여개의 전체 ORF 중 positive data와 같은 개수로 랜덤(random)으로 선정하여 실험을 하였다. 예를 들어 G1기에 대한 실험에서 사용된 데이터는 G1기에 관련된 것으로 알려진 ORF에 대한 데이터 257개, 나머지 ORF중 랜덤하게 선정된 데이터 257개, 총 514개가 사용되는 것이다.

표 1 실험에 사용된 전체 데이터 집합의 개수

	positive	negative	total
G1	257	257	514
S	69	69	138
S/G2	113	113	226
G2/M	182	182	364
M/G1	95	95	190

한편 인공신경망을 통한 학습에서 은닉 뉴런(hidden neuron) 수는 (클래스 수 + 모티프의 수)/2 로 결정하였으며, 학습 속도(learning rate)는 0.3, 모멘텀(momentum)은 0.2로 하여 학습하였다. 그리고 전체 데이터들은 5-fold cross validation을 통하여 학습되고 평가되었다.

3.2 실험 결과

실험을 통해 얻어진 스코어 값은 그림 2의 그래프를 통해서 나타난다. 각각의 단계에서 높은 점수를 가지는 것이 중요한 역할을 수행하는 모티프인 것으로 추정해볼 수 있다. 각 단계에서 높은 점수를 가지는 모티프는 표 2에 차례로 나타나 있다.

G1기의 경우 MCB 모티프가 다른 모티프들에 비해 월등히 높은 값을 가진다. MCB 모티프는 실제로 세포 주기 상에서 MBF라는 단백질과 결합하여 DNA의 전사 과정에 핵심적인 역할을 수행하는 것으로 밝혀져있다. 특히 세포 주기 상에서도 G1기에 MCB가 중요한 기능을 수행한다고 알려져 있다[2].

또한 S기에 가장 높은 점수를 가지는 SCB 모티프 역

시 SBF 단백질과 결합하는 부위로 세포 주기에서 중요한 역할을 수행하는 것으로 알려져 있다[2]. 그 외의 SFF, STRE, ECB, MET31-32, MCM1, ABF1, RAP1 등 각 단계에서 비교적 높은 값을 가지는 모티프들도 그 단계는 명확하지 않지만 세포 주기와 관련이 있는 것으로 추정되고 있는 모티프들이다[2, 3]. 이와 같은 기존의 논문 등에서 밝혀진 사실들을 생각해볼 때, 본 논문에서 사용된 방법이 실제 사실과도 부합된다고 볼 수 있다.

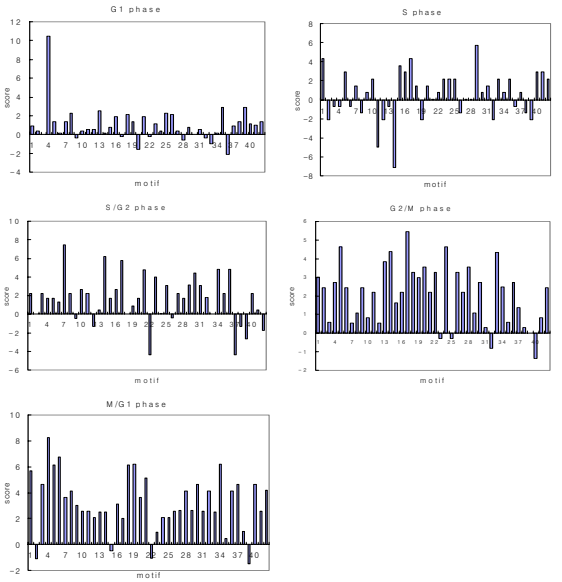

그림 2. 세포 주기 상의 각 단계에서 인공신경망 실험과 점수 계산을 통해 얻은 각 모티프의 중요도 점수 그래프

표 2. 인공신경망을 통해 결정된 세포 주기상의 각 단계별 모티프 중요도. 숫자는 식 (1)에 의해 결정된 점수값을 나타낸다. 가장 점수가 높은 것에서부터 차례로 15개의 모티프까지만 표에 보여주었다.

G1		S		S/G2		G2/M		M/G1	
MCB	10.47	SCB	5.75	AFT1	7.46	Leu3	5.47	MCB	8.29
SFF'	2.91	Leu3	4.32	ABF1	6.16	REB1	4.65	MIG1	6.75
ALPHA1'	2.90	RAP1	4.32	Leu3	5.72	HAP234	4.64	MET31-32	6.21
HSE	2.51	ATRepeat	3.60	BAS1	4.84	ABF1	4.37	SFF	6.18
STRE'	2.32	HAP234	2.89	SFF	4.83	MCM1	4.36	HAP234	6.17
STRE	2.31	GAL	2.89	PAC	4.78	HSE	3.83	LYS14	6.17
LYS14	2.12	ALPHA1	2.89	Gcr1	4.38	Yap1	3.56	RAP1	5.69
ECB	2.12	ALPHA2'	2.89	PHO	3.99	OAF1	3.55	PAC	5.16
GAL	1.94	STRE	2.18	SCB	3.10	PDR	3.29	Ume6(URS1)	4.66
PAC	1.93	CSRE	2.18	zap1	3.08	LYS14	3.29	ALPHA1	4.66
AFT1	1.36	MCM1	2.18	STRE	3.06	ECB	3.27	GCN4	4.65
HAP234	1.36	SFF'	2.18	CSRE	2.67	RAP1	3.01	Gcr1	4.63
SWI5	1.35	PHO	2.18	GAL	2.63	MET31-32	3.00	ALPHA2	4.19
MET31-32	1.35	REB1	2.18	SFF'	2.21	BAS1	2.74	Yap1	4.15
ALPHA2	1.35	ALPHA2	2.18	GCN4	2.20	MCB	2.73	MCM1'	4.13

물론 기존의 연구 결과에서는 세포주기와의 연관성이 알려지지 않은 모티프가 본 논문의 연구에서는 높은 결과값을 가지는 경우도 있다. 이러한 모티프들은 실제로 세포 주기에서 중요하게 작용하지만 아직까지 확인되지 않은 것일 가능성도 있다.

또한 전체 모티프 중 하나의 모티프 정보라도 없는 경우 그 정도의 차이는 있지만, 전체적으로 분류의 에러율이 높아지는 것을 볼 수 있다. 이는 모티프들이 단지 한두 개의 모티프에 의해서만 유전자 발현이 조절되는 것이 아니라, 결국에는 여러 개의 모티프 조합에 의해 유전자 발현이 조절된다는 사실도 이를 통해 유추해볼 수 있다.

4. 결론

본 논문에서는 인공신경망을 이용하여, 세포 주기 상의 각 단계에서 유전자 발현에 중요한 역할을 수행하는 모티프를 확인해보았다. 인공신경망을 이용한 학습 방법을 통해 정확도에 크게 영향을 미치는 모티프를 확인해 볼 수 있었다. 물론 이번 실험에서 확인된 모티프 계산값이 실제 세포 주기 상에서의 중요도를 정확히 표현하는 것은 아닐 수도 있다. 하지만, 많은 모티프 중에서 특별히 더 의미있는 모티프들을 찾아낼 수 있다는 점에서 유용성이 있다고 하겠다.

이번 논문에서는 단순히 하나의 모티프만에 대하여, 세포주기 상의 유전자 발현에 대한 중요도를 계산하여보았다. 향후 이 방법을 응용하면, 하나의 모티프만이 아니라 두 개, 혹은 그 이상의 모티프 조합에 의한 유전자 발현에 미치는 영향력을 밝혀내는 것도 가능할 것이다.

감사의 글

이 논문은 과학기술부의 국가지정연구실 사업(NRL)과 Systems Biology 사업(M10309000002-03B5000-00110)에 의하여 지원되었음.

참고 문헌

[1] Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein and Bruce Futcher, Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization, *Molecular Biology of the Cell,* Vol. 9:3273-3297, 1998.

[2] Raymond J. Cho, Michael J. Campbell, Elizabeth A. Winzeler, Lars Steinmetz, Andrew Conway, Lisa Wodicka, Tyra G. Wolfsberg, Andrei E. Gabrielian, David Landsman, David J. Lockhart and Ronald W. Davis, A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle, *Molecular Cell,* Vol. 2:65-73, 1998.

[3] Tu Minh Phuong, Doheon Lee and Kwang Hyoung Lee, Regression Tree For Regulatory Element Identification, *Bioinformatics,* Vol. 20(5): 750-757, 2004.

[4] Manuel Middendorf, Anshul Kundafe, Chris Wiggins Yoav Freund and Christina Leslie, Predicting Genetic Regulatory Response Using Classification, *Bioinformatics,* Vol. 20 Suppl 1: i232-i240, 2004.

[5] Yitzhak Pilpel, Priya Sudarsanam and George M. Church, Identifying Regulatory Networks by Combinatorial Analysis of Promoter Elements, *Nature genetics,* Vol. 29: 153-159, 2001.