서울시립대학교
UNIVERSITY OF SEOUL

서울시립대학교 자연과학대학 수학과

산업수학 연구인턴십

2020.09.18

고 지 형

2020 IEEE International Conference on Big Data and Smart Computing (BigComp)

Gene Expression Prediction using Stacked Temporal Convolutional Network

Imam Mustafa Kamal
Big Data Department
Pusan National University
Busan, South Korea
imamkamal@pusan.ac.kr

Nur Ahmad Wahid
Big Data Department
Pusan National University
Busan, South Korea
nawa410@pusan.ac.kr

Hyerim Bae*
Industrial Engineering Department
Pusan National University
Busan, South Korea
hrbae@pusan.ac.kr

***Abstract*— Predicting gene expression is one of the important tasks in molecular biology and genetics study. Studying the complex combinatorial code of gene expression could lead to a better understanding of gene regulation pattern i.e., how a gene increase or decrease specific gene products (protein and RNA) through translations. Such a pattern could be useful to study the origins of cancer, developing drugs for a certain disease, etc. In this study, we proposed to transform the Histone Modification data into one-dimensional space, and we predicted the gene expression by using Temporal Convolutional Networks. Previous studies proposed several methods ranging from classical machine learning approach (e.g., Support Vector Machine and Logistic Regression), as well as the most recent machine learning techniques (e.g., DeepChrome and DeepNN). Experiment results reveal that our approach is superior in terms of AUC score, accuracy, precision, recall, f-score, and specificity against the state-of-the-art-method, and only slightly worst in terms of precision and specificity against Support Vector Machine.**

Keywords — histone modifications, gene expression, classification, deep learning, temporal convolutional network

I. Introduction

Gene regulation is the process to control the expression gene transcription becomes possible to be carried on. Such a mechanism forms a kind of gene regulation. A study by [1] reveals evidence that specific histone modifications profiles may lead to cancer. However, epigenetic changes (e.g., histone modifications) are reversible. Thus, it leads to the importance of histone modification's role in developing drugs that can reverse such epigenetic change (e.g., for cancer treatments). For that reason, the gene expression prediction task becomes a hot research topic in the area of bioinformatics and molecular studies. Thus, several methods have been introduced to solve the problem.

A gene expression data classification using Support Vector Machine has been introduced [2]. A penalized logistic regression with adaptive LASSO for gene expression data for cancer classification task has been introduced [3]. The first research of gene expression prediction harnessing deep learning was DeepChrome [4] using Convolutional Neural Network (CNN) architecture. A deep neural network architecture to solve the gene expression prediction task also being introduced [5]. However, none of the approaches above consider the input being processed in a time-series manner.

In this paper, we propose a deep learning approach for

Main Points

유전자 발현 여부를 딥러닝 모델로 예측

히스톤 변형(histone modification) 데이터

TCN (Temporal Convolutional Network) 모델

클래식하게 활용되는 모델(SVM 등)에 비해 좋은 성능

딥러닝을 활용한 유전자 발현 예측

서울시립대학교 UNIVERSITY OF SEOUL

유전자 발현(Gene Expression)

세포 내부의 DNA는 유전자(gene)를 지님
유전자는 단백질(protein)을 암호화하고 있음
유전자에 의해 생성된 단백질은 세포의 기능을 결정
'유전자 발현' : DNA가 최종 생산물인 단백질을 생성하는 과정

1. RNA 중합효소(polymerase)가 오페론(operon) 내부 프로모터에 접착
2. mRNA 구축을 위해 연관된 유전자에게 전사(transcription)
3. mRNA는 번역(translation)을 위해 리보솜에 접근
4. 단백질 생성

*오페론: 조절유전자, 작동유전자, 프로모터, 구조유전자를 포함한 효소합성에 관여하는 일련의 DNA로 구성. 연관된 유전자들을 하나의 전사단위로 묶음으로써 연관된 유전자들을 통일적으로 조절

유전자 조절(Gene Regulation)

유전자의 발현을 조절하여 유전자가 생성하는 단백질의 양을 조절하는 것
세포 내에서 각 유전자별 발현되는 양을 조절하여 세포의 기능 결정함

히스톤 변형(Histone Modification)

후성유전학(epigenetics)

- DNA 염기서열의 변화 없이도 유전자 발현 패턴 및 유전자 발현 활성이 변화되고, 이것이 다음 세대로 유전되는 현상을 연구하는 학문
- DNA 메틸화(methylation) 연구가 가장 잘 알려져 있음

메틸화는 전사인자의 인식을 방해하며 일단 DNA가 메틸화되면 이 부위에서 '메틸기 결합(Methyl-Binding Domain, MBD) 단백질'이 유도된다. …(중략)… 유전자의 메틸화는 메틸기 전달효소의 활성보다 히스톤을 변형(histone modification) 또는 염색질 리모델링(chromatin remodeling)에 의해 전사과정을 억제하고 있다. 출처: https://blog.naver.com/hyouncho2/60090909848

히스톤 변형

- 히스톤 단백질에 대한 post-translational modification(PTM)
- 염색질의 구조를 변경하거나 히스톤 개질제를 모집함으로써 유전자 발현에 영향을 줄 수 있음
- 특정 히스톤 변형은 암을 유발하기도 하지만, 이러한 후성유전학적 변화는 가역적(reversible)
- 히스톤 변형 이해 => 후성유전학적 질병 치료 가능성

유전자 발현 예측 모델링

I. 데이터

학습 데이터: 히스톤 변형(histone modification) 데이터

- 각 유전자의 염기쌍을 N개의 구간으로 분할
- 각 HM(히스톤 변형)마다 TSS(Transcription Start Site) 추출
- 후성유전(Epigenetic) Feature 행렬 생성
- '유전자가 발현되었다(1)'/'되지않았다(0)'에 대한 레이블링
- 만들어진 행렬을 1차원 벡터로 stretch

*TSS: 중합효소(polymerase)가 붙으며 전사가 시작되는 위치

TABLE I. MULTI-DIMENSIONAL DATA OF GENE EXPRESSION

Row	*GeneID*	*HM1*	*HM2*	*HM3*	*HM4*	*HM5*	*Label*
1	1	0	7	0	5	2	1 (on)
2	1	0	0	1	7	0	1 (on)
…	…	…	…	…	…	…	…
100	1	1	2	0	8	9	1 (on)
1	2	0	1	4	0	0	0 (off)
2	2	0	1	8	1	3	0 (off)
…	…	…	…	…	…	…	…
100	2	4	0	1	3	7	0 (off)

TABLE II. ONE-DIMENSIONAL DATA OF GENE EXPRESSION

Row	*GeneID*	*1*	*2*	*…*	*…*	*…*	*499*	*500*	*Label*
1	1	0	7	…	…	…	8	9	1 (on)
1	2	2	0	…	…	…	3	7	0 (off)

II. 모델

TCN(Temporal Convolutional Network)

- CNN의 일종
- Causal Convolution 연산: 시계열성 반영

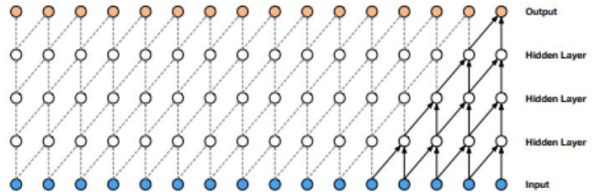
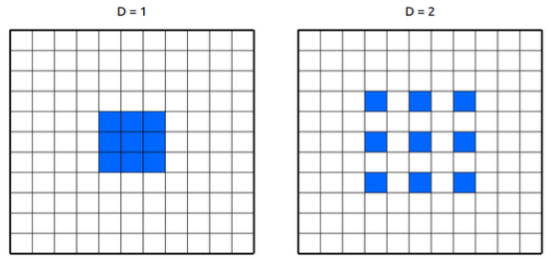


Figure 2: Visualization of a stack of causal convolutional layers.

Causal Convolution

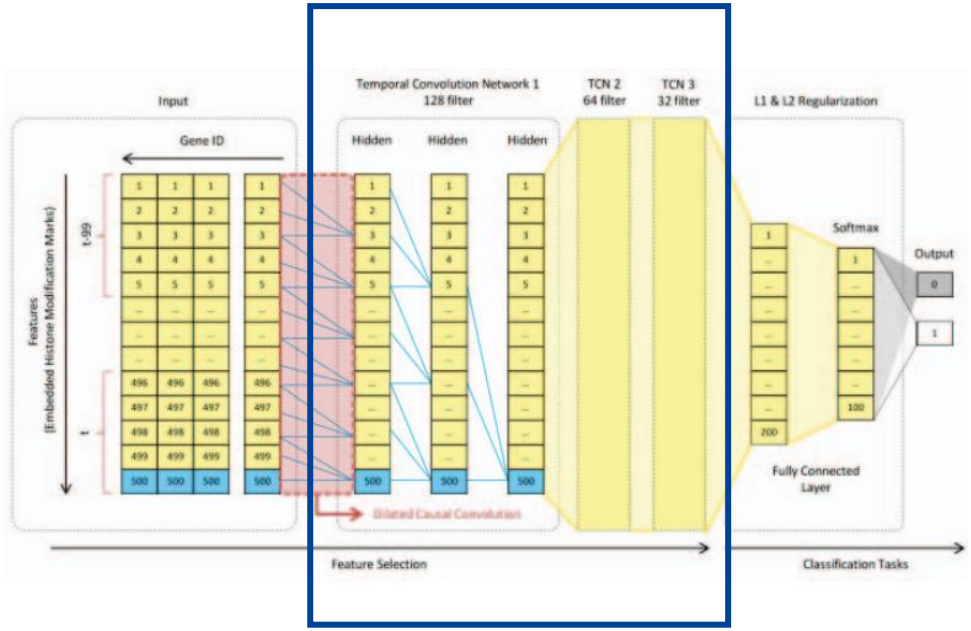
- Dilated Convolution 연산: 큰 데이터 필드에서도 학습 가능, 적은 메모리 비용



Dilated Convolution

유전자 발현 예측 모델링

III. 모델링 결과



- stretch된 1차원 인풋
- TCN 레이어가 Stack된 형태의 은닉층

IV. 학습 결과

- 클래식한 방식의 SVM, LR 모델보다 우월한 성능

TABLE III. EXPERIMENT RESULTS

METHODS	*AUC*	*ACC*	*PRE*	*REC*	*F*	*SPC*
TCN	**0.835000**	**0.837064**	0.758107	**0.825231**	**0.790247**	0.8440713
DeepChrome	0.828000	0.831683	0.752941	0.814815	0.782657	0.8416724
DeepNN	0.817000	0.824150	0.750412	0.789931	0.769665	0.8444140
SVM	0.826000	0.834266	**0.767897**	0.794560	0.781001	**0.8577793**
LR	0.825000	0.832329	0.763758	0.795139	0.779132	0.8543523

$$A = \int_{\infty}^{-\infty} TPR(T)\,FPR'(T)\;dt$$

A : Area under the ROC curve
$TPR(T)$: True positive rate given threshold T
$FPR(T)$: False positive rate given threshold T

$$ACC = \frac{TP+TN}{TP+FN+FP+TN} \quad \text{정확도}$$

$$REC = \frac{TP}{TP+FN} \quad \text{재현율}$$

$$F = 2\times\frac{REC\times PRE}{REC+PRE} \quad \text{F1-스코어}$$

$$PRE = \frac{TP}{TP+FP} \quad \text{정밀도}$$

$$SPC = \frac{TN}{TP+FN}$$

raw 데이터 출처: the Tampere University of Technology (published in Kaggle)
https://www.kaggle.com/c/gene-expression-prediction

감사합니다