마이크로RNA와 유전자 연관성 예측 딥러닝 모델 개발 연구

윤승원, 황인우, 김재인, 이규철*
충남대학교 컴퓨터공학과
[yoonenoch11,allhpy35,jaeinnn21]@gmail.com, *kclee@cnu.ac.kr

A Study on developing a deep learning model for predicting micro-RNA and gene association

Seung-Won Yoon, In-Woo Hwang, Jae-In Kim, Kyu-Chul Lee*
ChungNam National University

요 약

마이크로RNA는 인체에서 유전자가 단백질로 발현되는 정도를 조절하는 중요한 역할을 하는 RNA이다. 마이크로 RNA와 관련 있는 유전자를 찾는 연구는 복잡한 마이크로RNA의 메커니즘을 파악하는데 큰 도움이 된다. 본 연구는 마이크로 RNA와 유전자의 연관성을 예측하는 딥러닝 모델을 개발하는 연구이다. 논문에서는 최적의 파라메터 및 딥러닝 성능을 제시하며 해당 모델을 활용하여 다양한 실험을 진행하였다. 높은 성능을 보이는 학습된 모델을 통하여 모델이 처음 접하는 마이크로RNA-유전자 쌍의 연관성 정도를 스코어로 제시하며, 췌장암과 관련있는 유전자와 연관성 높은 마이크로RNA를 순위로 제시한다.

Ⅰ. 서 론

DNA의 유전정보는 RNA로 만들어지며 단백질을 합성하여 유전자가 발현된다. 대부분의 질병은 관련 단백질의 비정상적인 생성 및 과다 발현에 의해 발병된다. 마이크로RNA는 인체에서 유전자가 단백질로 발현되는 정도를 조절하는 중요한 역할을 하는 RNA이다. 마이크로RNA는 21-25개의 뉴클레오타이드(nucleotide)로 이루어진 단일 염기가닥의 리보핵산(small-RNA)이다. 마이크로RNA는 생명현상을 구동하는 단백질을 만드는 주형이 되는 메신저RNA를 분해함으로써 세포의 증식, 분화, 사멸을 조절한다. 즉, 질병을 일으킬 수 있는 단백질 생성을 차단 할 수 있는 기능을 가진다. 마이크로 RNA는 최근 질환의 발병원인과 밀접한 관련이 있다는 것이 밝혀짐에 따라 생물학, 약학 분야에서 매우 활발히 연구되고 있다[1].

마이크로RNA는 매우 복잡한 메커니즘을 가지기에, 마이크로RNA와 관련한 유전자를 정의하고 식별하는 연구는 마이크로RNA의 역할과 마이크로RNA의 메커니즘을 이해하는 것에 많은 도움을 줄 수 있다. 마이크로RNA와 관련한 유전자를 정의하는 연구는 마이크로RNA와 관련한 질병을 예측하는 연구에 상당한 기여를 할 수 있으며, 해당 마이크로RNA를 통한 질병치료 가능성을 알 수 있다. 또한 마이크로RNA와 관련한 약물을 예측하는 연구에도 기여할 수 있으며, 약물을 통한 마이크로RNA의 활성정도를 조절할 수 있다[2].

마이크로RNA와 관련한 유전자를 알기 위해 전통적인 생물학적 약학적 실험을 하는 것은 마이크로RNA 자체를 구하기도 어려울 뿐더러 매우 복잡한 관계성을 띄기에 수많은 시간과 비용을 필요로 한다. 그리하여 직접 실험이 아닌 모델 등을 활용한 계산적 방법(computational method)으로 연관성을 예측하는 연구들이 존재한다. 그 중에서도 본 논문은 딥러닝 모델을 생성하여 마이크로RNA와 유전자의 연관성을 예측하는 연구이다.

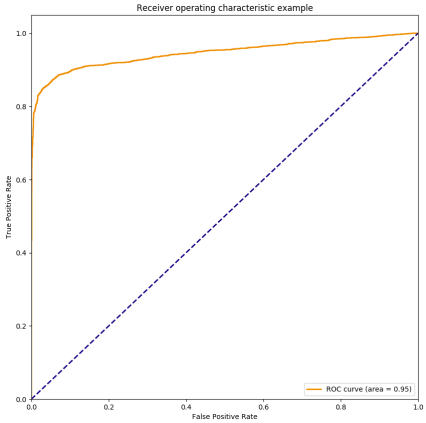
본 논문의 구성은 2장에서는 마이크로RNA 데이터와 유전자의 관련성을 예측하는 딥러닝 모델에 대하여 설명하고, 해당 모델의 성능과 예측실험 결과를 제시할 것이며 3장에서는 결론 및 향후 연구를 제시한다.

Ⅱ. 본론

마이크로RNA, 유전자, 질병 등 여러 상관관계를 찾는 연구는 기존에도 존재하였다. 하지만, 기존 대부분의 연구들은 규칙기반(Rule based) 방법과 기계학습(Machine learning) 방법으로 연관성을 예측하였다. 하지만 주어진 입력에 대해서만 결과를 도출하는 규칙기반학습과 기계학습은 일일이 인간이 특징(feature)를 지정해야 하기에, 데이터 자체가 복잡한 도메인인 유전자단의 연구에서는 한계점이 존재한다. 최근 이러한 한계점을 극복하기 위해 유전자 도메인의 연관성 예측 딥러닝 연구가 이루어지고 있다[3]. 본 연구는 마이크로RNA와 유전자의 연관성을 예측하는 딥러닝 모델을 개발하는 연구이다.

다른 학습모델과 마찬가지로 딥러닝 모델에서 데이터는 굉장히 중요한 역할을 한다. 마이크로RNA와 유전자의 특징들을 정교하고 정확하게 추출하는 것이 해당 관련 분야의 주된 연구가 될 정도로 딥러닝 모델에 정확한 데이터를 넣어 주는 것이 중요하다. DNA는 아데닌(A), 구아닌(G), 시토신(C), 티민(T)의 염기서열로 구성되어있으며, RNA는 아데닌(A), 구아닌(G), 시토신(C), 우라실(U) 이렇게 4가지 염기서열로 구성되어있다. 이렇듯 유전자는 일종의 시퀀스를 갖기에 유전자 관련 딥러닝 연구에서는 주로 시퀀스(sequence) 특징만을 추출하여 학습데이터로 활용한다. 하지만 이러한 시퀀스 이외에 각각의 유사도를 기반으로 네트워크를 생성하여 해당 네트워크의 지형적(geometric) 특징을 추출하는 데이터가 존재한다. 본 연구에서는 주로 활용되는 시퀀스 특징데이터 뿐만 아니라 지형적 특징데이터를 활용하여 딥러닝 모델을 생성하였다. 활용된 데이터는 각 특징 별로 128차원으로 되어 있으며, 레이블링(labeling)은 유클리디언(euclidian)유사도 및 코사인(cosine) 유사도를 기반으로 이루어졌다. 해당 마이크로RNA와 유전자의 유사도 거리가 일정 거리 이상이면 0(연관없음), 일정거리 이하면 1(연관있음)으로 레이블링 하였다.

본 연구는 마이크로RNA와 유전자의 연관성을 예측하기 위해 LSTM (Long-Short Term Memory) 딥러닝 모델을 개발하였다. LSTM 모델은



[그림1]. 본 연구 딥러닝 모델의 ROC 커브

RNN(Recurrent Neural Network) 계열의 딥러닝 모델이다. 유전자들은 시계열 특징을 갖기에 RNN 계열의 딥러닝 모델을 개발하였다. LSTM모델은 RNN의 기울기 소실 문제를 어느 정도 해결한 딥러닝 모델이다.

본 연구의 딥러닝 모델 입력(Input)데이터는 각 128차원 RNA와 유전자의 특징 데이터들이며 생성한 딥러닝 모델은 해당 RNA와 유전자가 연관이 있는 것인지(label:1) 없는 것(label:0)인지를 예측하는 것이다.

본 연구의 딥러닝 모델은 LSTM 레이어(layer) 3으로 지정하였으며, 손실함수(loss function)는 예측 값을 0과 1 사이 값으로 변환하여 손실정도를 계산하는 Cross Entropy를 활용하였다. Cross Entropy는 주로 분류문제에 많이 활용된다. 최적화 함수(Optimization function)는 아담 최적화 함수를 활용하였다. 아담 최적화 함수는 학습에서의 스텝사이즈가 기울기의 스케일링에 영향을 받지 않는다는 특징이 있다. 기울기가 커져도 스텝사이즈는 한정되어 있기에 어떠한 목적함수(objective function)를 사용하더라도 안정적인 최적화를 위한 하강이 가능하다는 특징을 가진다.

모델 구축 시험에 활용된 전체 데이터는 마이크로RNA와 유전자 쌍으로 이루어져있으며 총 31,080개이다. 학습데이터와 성능측정을 위한 테스트 데이터 개수는 8:2 비율로 랜덤하게 지정하였다. 배치(batch) 사이즈는 64로 지정하였으며, 학습률(learning rate)는 0.001로 지정하였다. 본 연구의 모델 성능은 **[그림1]**에서 제시한 바와 같이 ROC 커브 0.95의 성능을 보인다. ROC 커브는 대표적인 딥러닝 모델의 성능을 나타내는 수치이며 최고 값은 1이다. 모든 실험은 Intel(R) Core(TM) i7-7700 CPU @ 3.60GHz, 32GB RAM, GeForce GTX 1080 Ti GPU 환경에서 진행하였다.

우수한 성능을 보이는 본 연구의 딥러닝 모델을 통하여 학습과 테스트에 활용되지 않은 마이크로RNA와 유전자 1,000쌍의 연관성 정도를 예측하는 실험을 진행하였다. 본 연구의 딥러닝 모델은 연관성 정도를 스코어로 나타낼 수 있으며, 그 결과를 연관성스코어 기준 내림차순으로 나열하여 Top-10으로 제시한 것이 **[그림2]**이다. 해당 실험의 연관성스코어가 높은 수치를 보이지 않는 이유는 이미 학습과 테스트에 활용된 데이터를 배제하고 관련성이 그리 높지 않은 데이터를 통한 실험이기 때문이다.

본 연구팀은 해당 딥러닝 모델을 통해 추가적인 실험을 진행하였다. 인류 최악의 암으로 불리는 췌장암의 연관 유전자 중에는 MAPK1이 있다. 본 연구팀은 해당 MAPK1 유전자와 관련 있는 마이크로RNA를 예측하는 실험을 진행하였다. 해당 실험의 결과는 **[그림3]**과 같다. 본 연구의 모델은 췌장암과 관련 있는 마이크로RNA를 연관성 순위로 제시할 수 있으며, 췌장암 뿐 아니라 다른 질병과 관련 있는 마이크로RNA 및 유전자의 연관성 정도 예측할 수 있음을 제시한다.

순위	마이크로RNA	유전자	관련성스코어
1	hsa-let-7c-5p	ATG9A	34.8029137
2	hsa-miR-4257	RPL41	33.5258217
3	hsa-miR-6779-5p	STK38	33.2252464
4	hsa-miR-15a-5p	KIF5B	32.5363503
5	hsa-miR-4430	RACGAP1	32.3173752
6	hsa-miR-4668-5p	TMBIM6	31.5996399
7	hsa-miR-4257	SCAMP4	31.4043465
8	hsa-miR-302a-3p	SF3B3	30.9225254
9	hsa-miR-3689b-3p	SCAMP4	30.855587
10	hsa-let-7a-5p	ZNF566	30.7286434

[그림2]. 관련성 스코어 Top-10 결과

순위	유전자	마이크로RNA
1	MAPK1	hsa-miR-519d-3p
2	MAPK1	hsa-miR-20a-5p
3	MAPK1	hsa-miR-17-5p
4	MAPK1	hsa-miR-93-5p
5	MAPK1	hsa-miR-660-3p
6	MAPK1	hsa-miR-548aq-3p
7	MAPK1	hsa-miR-526b-3p
8	MAPK1	hsa-miR-6829-3p
9	MAPK1	hsa-miR-519b-3p
10	MAPK1	hsa-miR-20b-5p

[그림3]. MAPK1과 관련있는 마이크로RNA 예측 결과

Ⅲ. 결론

본 논문에서는 마이크로RNA와 유전자의 연관성을 예측하는 딥러닝 모델을 개발 및 추가적인 실험 결과를 제시하였다. 본 연구의 딥러닝 모델을 통해서 여러 마이크로RNA와 유전자의 연관성 정도를 스코어로 예측할 수 있기에, 많은 질병과 관련 있는 유전자 등을 예측할 수 있다. 하지만, 본 연구의 모델이 예측한 연관성 있는 쌍들 검증은 기존에 연관성이 알려진 쌍들만 가능한 한계가 있다.

향후에는 약학대학 연구실과 협업을 통해 기존 연관성이 알려진 마이크로RNA와 유전자 데이터 이외에 본 연구팀에서 예측한 결과들을 검증하는 실험들을 추가적으로 진행할 것이다.

ACKNOWLEDGMENT

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2020-0-01441, 인공지능융합연구센터지원(충남대학교))

참 고 문 헌

[1] Fu, Laiyi, and Qinke Peng. "A deep ensemble model to predict miRNA-disease association." Scientific reports 7.1 (2017): 1-13.

[2] Deepthi, K., and A. S. Jereesh. "An Ensemble Approach Based on Multi-Source Information to Predict Drug-MiRNA Associations via Convolutional Neural Networks." IEEE Access 9 (2021): 38331-38341.

[3] Pla, Albert, Xiangfu Zhong, and Simon Rayner. "miRAW: A deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts." plos computational biology 14.7 (2018): e1006185.