https://doi.org/10.7840/kics.2023.48.10.1289

Distance metric 기반 detailed Negative set을 사용한 원형 RNA와 질병 연관성 예측

황 인 우•, 윤 승 원*, 김 재 인*, 이 규 철°

CircRNA-Disease Association Prediction Using Detailed Negative Set on Distance Metric-Based(DiNeg-CDA)

In-Woo Hwang•, Seung-Won Yoon*, Jaein-Kim*, Kyu-Chul Lee°

요 약

CircRNA는 인간에게 치명적인 질병인 알츠하이머나 심혈관 질환과 관련되어 있다. 생물학적 실험을 통해 위험한 질병과 관련된 CircRNA를 밝혀내기 위해 많은 시간과 돈이 소모된다. 시간 또는 자원을 절약하는 효율적인 접근법 중 하나가 딥러닝을 활용하는 것이다. 본 연구는 Distance metric을 활용하여 벡터 공간에 존재하는 Negative set을 두 가지 기준을 가지고 Positive와 유사한 데이터를 삭제하여 Random negative set 보다 정교한 Negative set 구축 방법을 제안한다. 비교 실험을 통해 정교한 Negative set이 Random set보다 좋은 성능을 보였다. 제시한 모델의 DiNeg-CDA의 모델 성능은 AUC-ROC curve 0.89로 측정되었다.

키워드 : 원형RNA, 네거티브 셋, 원형RNA-질병 연관성 예측, 특징 추출, 거리 함수

Key Words : CricRNA, Negative set, circRNA-disease association prediction, Feature extraction, distance metric

ABSTRACT

ScircRNAs have been implicated in Alzheimer's and cardiovascular diseases, which are fatal diseases in humans. A lot of time and money are spent trying to identify circRNAs associated with dangerous diseases through biological experiments. One of the efficient approaches to save time or resources is to utilize deep learning. This study proposes a more sophisticated negative set construction method than the random negative set by deleting data similar to the positive with two criteria for the negative set existing in the vector space using distance metric. Through comparison experiments, the sophisticated Negative set performed better than the Random set. The model performance of DiNeg-CDA of the proposed model was measured with an AUC-ROC curve of 0.89.

※ 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2021R1F1A1061513)
◆ First Author : Chungnam National University, allhpy35@gmail.com, 학생회원
° Corresponding Author : Chungnam National University, kclee@cnu.ac.kr, 정회원
* Chungnam National University, yoonenoch11@gmail.com, 학생회원; National University, jaeinnn21@gmail.com, 학생회원
논문번호 : 202307-017-C-RU, Received July 23, 2023; Revised July 27, 2023; Accepted July 27, 2023

Ⅰ. 서 론

1.1 생명활동과 유전자 발현

우리 몸에서 일어나는 생명활동은 외부 환경으로부터 오는 자극에 대응하면서 생명을 유지한다. 우리 몸은 외부로부터 오는 모든 외부 자극들은 생명활동을 이어나가는데 방해되는 요소로 인식한다. 외부 자극과 같은 방해 요소들은 세포가 대응하게 된다.

우리 몸은 단백질 양을 조절하여 외부 자극을 대응한다. 왜냐하면 세포의 기능은 세포에 존재하는 단백질의 양에 따라 결정되기 때문이다. 따라서 세포 내부의 단백질의 양에 따라 세포가 수행하는 일이 달라지며 이 활동은 외부 자극에 대응하게 된다. 몸에서 필요한 단백질을 생성하기 위해 수행되는 과정을 유전자 발현(Gene expression)이라 한다.

신기하게도 우리 몸은 외부의 자극을 과하게 대응하지 않는다. 이 말은 각 기관의 세포 내부에서 필요한 양만큼만 단백질을 조절하여 생성한다는 의미이다. 이렇게 단백질의 양을 조절하는 것을 유전자 발현의 조절(Gene regulation)이라 한다.

유전자 발현의 조절을 담당하는 RNA가 COVID-19 시대에 각광받은 마이크로RNA(microRNA, miRNA,)이다. miRNA는 약 22개의 뉴클레오타이드(nucleotide)로 구성된 짧은 non-coding RNA이다. miRNA는 인간의 유전자 60%를 조절한다[1]. 최근 miRNA와 결합하여, 유전자 발현을 조절할 수 있는 원형 RNA(circular RNA, circRNA)에 대해 주목받고 있다[2].

기존의 선형 RNA(linear RNA)와는 달리 원형 RNA는 closed-loop 구조로 되어있는 단일 가닥 RNA(single-standard RNA)의 한 종류이다. 약 30년 전 mRNA의 전구체로부터 백 스플라이싱(Back-splicing) 과정을 통해 생성된다[2,3]. 하지만 발견 이후에 소수의 원형 RNA들만 발견되어 기능이 없는 부산물로 여겨져 왔다. 최근 차세대 염기서열 분석(Next-Generation Sequencing, NGS) 기술로 세포의 조직에 따른 발현량이 확인되면서 더욱 주목 받게 되었다.

1.1.1 원형 RNA와 질병과의 관계

원형 RNA의 돌연변이 또는 비정상적인 활동들이 질병을 일으킬 수 있다[3,4]. 아래의 예시처럼 실제 치명적인 질병들과 관련 있어서 문제가 된다. 태아 발달 10주와, 20주 때 심장에서 원형 RNA가 발견되었고, 원형 RNA가 심혈관계에 밀접한 관련 있다는 사실이 생물학적 실험을 통해 증명되었다[4]. 그리고 알츠하이머 환자의 뇌에서 원형 RNA가 발견되면서 치매의 중증도와

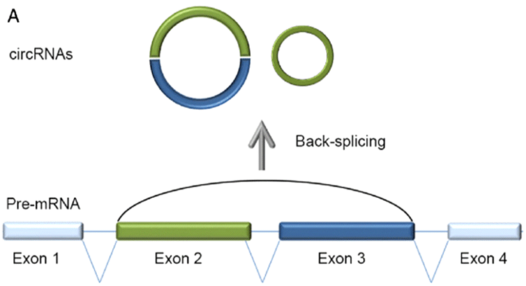

그림 1. 원형 RNA의 closed-loop 구조
Fig. 1. CircRNA of Closed-loop structure

관련 있다[3].

그림 1 원형 RNA은 Closed-loop 구조로 인해 우리 세포에서 외부의 환경 변화에 둔감하다. 이러한 안정적인 특성이 바이오마커로 선정되었을 때 큰 장점을 가지게 된다.

바이오마커(Biomarker)는 몸 안의 변화를 객관적으로 측정할 수 있는 지표를 말한다[5]. 일반적으로 바이오마커를 통해 병의 유무 진단, 치료 반응 예측할 수 있다. 실제 골육종 환자를 치료하기 위한 원형 RNA가 바이오마커로 사용되고 있다. 골육종뿐만 아니라 다른 원형 RNA와 질병의 연관성을 예측하는 연구는 관련된 질병을 치료할 수 있는 중요한 연구이며, 원형 RNA와 관련된 질병을 분석할 수 있는 다양한 시도들이 필요하다.

1.1.2 바이오 산업에서 딥러닝 도입의 장점

대표적인 바이오산업에는 신약 개발 및 의약품 개발이 있다. 질병을 치료하기 위한 의약품 개발 기간은 평균 15년 소모된다. 약 5,000개 ~ 1만여 개의 후보군 중에서 1개 만이 최종 신약 개발에 성공하게 된다[6].

지난 15년간 미국 제약사들은 신약 개발에 520조 원을 투입하였으며, 이는 항공산업에서의 5배에 해당하는 규모이다[6]. 다양한 산업에서 질병을 치료하기 위해 큰 비용과 시간을 투자하고 있다.

그림 2[6]에서 딥 러닝으로 기존의 개발과정을 대체함으로써 얻는 이점은 두 가지가 있다.

첫 번째 최대 10년 동안 했던 인간의 일을 딥 러닝을 통해 2년으로 단축할 수 있다. 이것은 기존에 소모되었던 시간 및 자원을 절약할 수 있는 효율적인 방법이다. 두 번째 기존의 생물학적 지식의 비중을 줄 일수 있다. 기존의 개발과정에서는 인간이 직접적으로 실험 환경을 통제했기 때문에 인간의 지식이 중요하게 적용되다. 하지만 딥 러닝 모델은 상대적으로 덜 중요하다. 왜냐하면 데이터에서 중요한 부분을 인간이 알려주지 않아도 딥 러닝 모델이 스스로 찾아서 학습하기 때문이다. 딥러

닝을 활용한 방법은 시간 및 자원을 절약할 수 있고, 기존 개발과정보다 효율적인 접근 방법이다.

1.1.3 CircRNA-disease association prediction 연구의 데이터 부족으로 인한 Positive, Negative set구축 방법

CircRNA-disease association prediction 연구에서 사용하는 데이터셋은 생물학자들이 정밀한 실험을 통해 밝혀낸 "질병A와 관련 있다"라고 정의된 데이터(Positive)이다. 이 데이터를 벡터 공간(Vector space)에 위치시켜 데이터 사이의 복잡한 관계를 딥 러닝 모델이 예측한다.

딥 러닝 모델이 복잡한 관계를 정확하게 예측하기 위해서는 적절한 벡터 공간에 데이터들을 위치시켜야 한다. 하지만 단순하게 Positive 데이터를 바로 벡터 공간에 위치 시킬 수 없다. Positive 데이터의 고유한 특징을 반영하여 벡터로 만드는 과정인 임베딩(Embedding)이 중요하다. 임베딩 과정을 거친 후 생성된 벡터 데이터(feature)를 활용하여, 딥러닝 모델이 학습할 수 있는 학습 데이터 셋(Train set)을 구축한다.

일반적으로 질병과 관련된 원형 RNA들의 관련성 예측 연구에서 학습데이터 셋의 구축 비율(Positive set와 Negative set의 비율)을 1:1로 사용한다. 하지만 Positive와 반대되는 관계인 "어떠한 질병과 관련이 없다"(Negative)라는 것은 생물학적으로 정의 할수 없다. 왜냐하면 영속적으로 불변하는 관계를 정의할 수 없기 때문이다. 그래서 기존 연구에서 Positive에 존재하지 않는 모든 관계를 Negative 후보군으로 정의하고, 같은 비율의 데이터를 랜덤으로 추출한다. 랜덤으로 선택하는 기존 연구들의 방법은 정교한 Negative set이라 할 수 없다. 원형 RNA와 질병의 관련성 예측 연구의 Positive 데이터가 다른 RNA(miRNA-disease, lncRNA-disease)를 사용한 연구들 보다1/3 정도로 작은 규모에 속한다. 그래서 주로 여러 Positive 데이터셋을 통합하여, 데이터를 확보한다.

Ⅱ. Distance metric 기반 detailed Negative set을 사용한 원형 RNA와 질병 연관성 예측(DiNeg-CDA)

본 연구에서는 세 가지 거리 기반 메소드 Mahalanobis distance(마할라노비스 거리), Cosine similarity(코사인 유사도), Euclidean distance(유클리드 거리)를 사용하여, 기존 연구들 보다 정교한 Negative set을 구축하는 방법을 제안한다. 그리고 기존 연구보다 약 10배 큰 규모의 데이터 셋을 구축하였고, 제안한 Negative set 구축 기법을 구축한 데이터 셋에 적용하였다. Negative 기법까지 적용한 데이터 셋은 기존 연구에 사용된 데이터 셋보다 규모도 크고 정교한 Negative를 가진 학습 데이터(Train data)이다. 구축한 학습 데이터를 활용하여, 원형 RNA와 질병 연관성을 예측할 수 있는 LSTM(Long Short-Term Memory) 예측 모델을 제안한다. LSTM 모델의 성능은 5 fold를 통해 제시하였다. 본 연구에서 제시한 모델의 최고 성능은 0.89를 달성하였다.

2.1 논문의 구성

본 논문의 3장에서는 관련 연구에서 제시한 원형 RNA와 질병의 각각의 embedding 방법과 예측 과정을 간략히 설명한다. 그리고 기존 연구의 문제점을 제시한다.

4장에서는 circRNA-disease association prediction 연구에서 구축하는 데이터셋 구축 시 문제점을 제시한다. 이 문제를 해결할 정교한 negative set의 필요성을 주장한다.

5장에서는 기존 연구보다 약 10배 큰 데이터셋을 구축하는 방법을 자세히 설명한다. 그리고 구축한 데이터셋을 활용하여 기존 원형 RNA를 embedding 한 벡터를 검증할 수 있는 연구를 소개한다. 그리고 그 방법을 통해 생성한 데이터셋의 positive, circRNA, disease 구축 과정 및 완성된 데이터의 수를 공개한다.

6장에서는 embedding된 데이터를 거리 기반 Negative set을 적용한 설명을 한다. 그리고 본 연구에서 사용한 LSTM 모델의 구조를 설명한다.

마지막 7장에서는 본 연구에서의 실험 성능과 향후 연구 및 결론을 제시한다.

Ⅲ. 관련 연구

일반적으로 원형 RNA는 염기서열 데이터를 사용하여 고유한 특징을 나타낸다. 하지만 질병의 특징 데이터

그림 2. 기존 개발과 딥 러닝을 활용한 개발 과정 비교
Fig. 2. Comparison of development process using existing development and deep learning

는 존재하지 않기 때문에 관련 연구들이 제시한 질병 특징 추출 기법이 조금씩 다르다.

3.1 CDASOR 연구

CDSOR[7]에서는 질병의 특징을 추출하기 위해 DO(Disease Ontology,)를 사용하였다. DO의 Categorical 한 구조를 이용하여, 그래프로 나타냈다. 각각의 노드가 질병 이름이 되며, 질병과 질병 사이의 관계(Edge)는 상위 질병 노드에 존재하면, 연결되는 구조이다. DO의 그래프 관계를 직렬화하여, GloVe를 통해 embedding이 진행된다.

CDSOR[7] 연구는 원형 RNA의 의미를 담기 위해 NLP(Natural Language processing) 분야에서 Corpus (말뭉치)를 생성하는 방법과 유사한 K-mer를 사용한다. K-mer를 통과한 각각의 염기서열들이 GloVe를 통해 임베딩 된다.

K-mer는 하나의 염기서열을 N개 만큼 잘라서 나타내는 방법이다. N을 일반적으로 length라 부르며, 그림 3의 length는 3이다. CDSOR[7]에서는 하나의 염기서열 (ATTHATTGTC)이 3-mer를 통해 ATT, TTG, TGA… 으로 표현된다.

그림 3. K-mer 생성 예시
Fig. 3. K-mer creation example

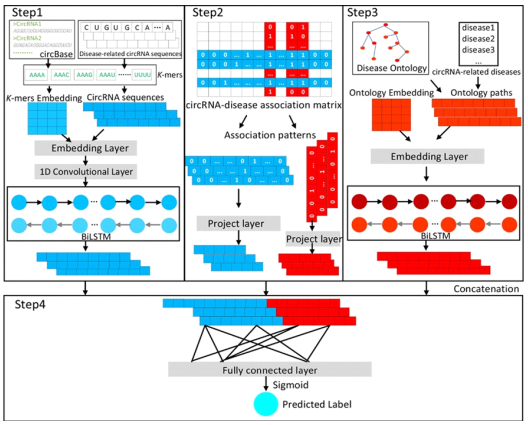

그림 4. CDSOR의 흐름도
Fig. 4. CDASOR workflow diagram

그림 4의 CDSOR[7] 흐름도를 각 Step을 기준으로 설명하겠다. Step 1: 원형 RNA를 k-mer 통해 나온 각각의 염기서열과 기존의 염기서열을 Embedding layer(Glove)를 통해 벡터화한 후 1D convolutional layer를 통해 특징을 추출한다.

Step 2: circRNA-disease 관계데이터(positive data)를 기준으로 하나의 원형 RNA와 관련된 질병의 관계가 존재하면 1 없으면 0으로 표시하여 인접 행렬(Adjacency matric)을 나타낸다. 생성한 인접 행렬을 원형 RNA와 질병으로 분리하고 project layer를 통해 연관성 패턴을 추출한다.

Step 3에서 DO와 positive data에서 추출한 질병 들을 Embedding layer를 통과시킨 후 BiLSTM(Bidirectional LSTM)에서 질병들의 관계를 학습한다. 이 부분은 Step 1에서 BiLSTM 부분과 동일한 과정이며, 원형 RNA들의 관계를 학습한다. 마지막으로 Step 4에서 Fully connected layer에서 질병과 원형 RNA 관계 벡터를 생성한다. 이 단계에서는 Step 1,2,3에서 표현된 각각의 원형 RNA, 질병 그리고 positive data가 통합된다. 그 후 관계가 반영된 벡터를 sigmoid 함수를 통과하여 확률 값으로 원형 RNA와 질병의 관계를 예측하게 된다.

3.1.1 IMS-CDA 연구

IMS-CDA[4] 연구에서는 원형 RNA의 염기서열 데이터를 사용하지 않고, 원형 RNA의 특징을 나타내는 것이 기존 연구와 다른 점이다. 원형 RNA 또는 질병의 기준으로 생성한 두 개의 인접 행렬을 통해 각각의 특징을 추출한다. 예를 들어서 원형 RNA의 기준의 인접 행렬은 1행에는 하나의 원형 RNA이며, 열은 중복되지 않은 모든 질병으로 구성되어 있다. 하나의 행에는 하나의 원형 RNA와 관련된 모든 질병의 관계가 정의된다. 생성한 인접 행렬을 통해 Similarity(유사도)를 계산하면 하나의 원형 RNA와 관련된 질병들의 유사한 정도를 표현할 수 있다.

위의 방법으로 원형 RNA와 질병 각각의 인접 행렬을 생성한 다음에 특징 추출이 진행된다. 원형 RNA의 특징은 Jaccard similarity를 통해 질병의 유사도를 나타낸다. 그 다음에는 GIP(Gaussian Interaction Profile kernel similarity)통해 원형 RNA 관점에서 질병의 유사도를 임베딩 하게 된다. 질병의 특징을 추출하기 위해 앞서 말한 인접 행렬과는 반대의 기준으로 인접 행렬을 생성한다. 생성한 인접 행렬에 Jaccard similarity 계산하여 유사도를 계산한다. 원형 RNA와 질병에 동일한 유사도 함수를 사용한 것은 동일한 관점으로 유사도를

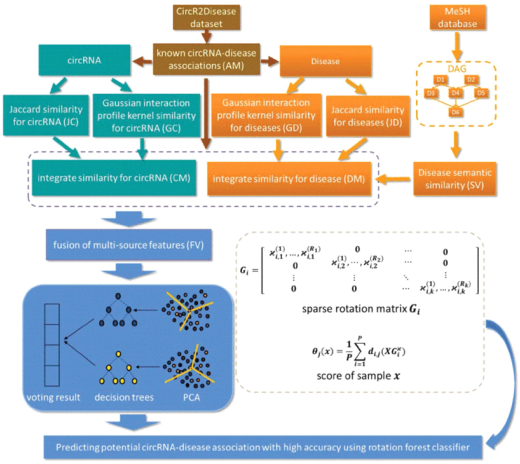

그림 5. IMS-CDA의 흐름도
Fig. 5. IMS-CDA workflow diagram

나타내기 위해서이다. 질병의 임베딩은 MeSH(미국 국립의학도서관 데이터베이스)의 Categorical 정보를 그래프로 생성하여 문서적인 질병의 정보를 추출한다. 하지만 Positive 데이터 셋에 존재하는 질병들이 MeSH에 있는 개수와 질병의 명칭을 밝히지 않았다. 그래서 어떤 질병의 문헌정보가 반영되었는지 알 수 없다.

IMS-CDA[4] 연구의 전체적인 구조는 그림 7이다. 임베딩된 각각의 원형 RNA와 질병 데이터를 Positive 관계에 맞게 매칭하여, Fusion of multi-source feature(FV)를 생성한다. 그 후 PCA 통해 차원을 축소시킨 후 의사결정 트리(Decision tree)를 통해 나온 결과를 Score로 변환하여 원형 RNA와 질병의 관계를 예측하게 된다. 그리고 모든 Negative 후보군에서 랜덤으로 추출하여 Negative set을 생성한다.

관련 연구(IMS-CDA[4], CDSOR[7])에서는 Negative set을 생성할 때 모든 Negative 후보군에서 랜덤으로 선택하여, 구축하였다. 하지만 랜덤으로 선택한 Negative set에는 딥 러닝 모델이 판단하기 어려운 데이터 즉 Positive 보다 더 유사한 Negative가 존재한다. 앞서 말한 데이터 때문에 모델의 성능은 낮아지며, 일반화를 방해 하기 때문에 정교한 Negative set이 필요하다.

Ⅳ. Detailed Negative set 구축을 위한 데이터 통합과 embedding 기법

4장에서는 먼저 원형 RNA와 질병의 연관성 예측 연구에서 주로 사용하는 데이터 확보 방법과 확보한 데이터 셋을 활용하여 학습 데이터 구축 시 문제점을 이야기한다. 그리고 앞서 문제들을 해결한 기존 연구들보다 큰 규모의 원형 RNA-질병 데이터 셋((Positive set) 구축 과정을 자세히 설명한다. 그리고 구축한 데이터 셋의 구성 요소(원형 RNA의 수, 질병의 수, Positive 수)를 명확히 공개하고, 각 원형 RNA와 질병의 embedding 과정을 설명한다.

4.1 통일되지 않은 원형RNA의 symbol 문제

원형 RNA와 질병 연관성 예측을 하는 연구는 주로 공개된 데이터셋(Open Dataset)을 사용한다. 그림 6의 CircR2Disease[9]는 circRNA-disease association prediction에서 대표적으로 사용되는 Open dataset(Positive set) 중 하나이다. Positive set은 원형 RNA와 관련된 질병 이름의 쌍으로 존재한다. 그림 6에서 circRNA Name 열(점선, 파란색)이 각각의 원형 RNA의 symbol를 의미한다. 그리고 Disease Name 열(실선, 주황색)이 질병의 명칭이다. 그리고 그림 6의 circRNA Name 열에 다른 원형 RNA들이 함께 표기되어 있다. Symbol이 다양한 이유는 데이터베이스마다 각기 다른 기준으로 원형 RNA를 표기하기 때문이다. 원형 RNA의 통일화되지 않은 symbol 때문에 positive set의 수를 결정짓기 어렵게 한다. Symbol의 통일화를 위해 원형 RNA 염기서열 데이터베이스 중 가장 큰 규모이며, 148,175개의 sequence를 보유하고 있으며, 다른 원형 RNA 데이터베이스에서 참조 이름으로 사용되고 있는 CircBase[10]를 사용하였다.

본 연구에서는 그림 7 CircBase[10]에 표기된 원형 RNA를 기준으로 중복되는 원형 RNA의 이름들을 하나로 통일하였다. 그리고 그림 7에는 두 형태의 원형 RNA가 존재한다. 둘의 차이점은 has_circ뒤에 숫자의 수(왼쪽 7개, 오른쪽 6개)가 다르다. 대부분의 데이터베이스

hsa_circ_102584/hsa_circ_0003146	chr19:48+	EHD2	http://www.circbase.	Systemic lupus erythematosus
hsa_circ_101471/hsa_circ_0034398	chr15:36+	C15orf41	http://www.circbase.	Systemic lupus erythematosus
hsa_circ_104807/hsa_circ_0001866	chr9:862-	UBQLN1	http://www.circbase.	Systemic lupus erythematosus
hsa_circ_102571/hsa_circ_0008615	chr19:45-	PPP1R13	http://www.circbase.	Systemic lupus erythematosus
hsa_circ_100226/hsa_circ_0005567	chr1:518-	EPS15	http://www.circbase.	Systemic lupus erythematosus
hsa_circ_100775/hsa_circ_0021549	chr11:30-	MPPED2	http://www.circbase.	Systemic lupus erythematosus
hsa_circ_101889/hsa_circ_0040705	chr16:84+	USP10	http://www.circbase.	Systemic lupus erythematosus
hsa_circ_0001649	chr6:146-	SHPRH	http://www.circbase.	Glioma
hsa_circ_0007534	chr17:61+	DDX42	http://www.circbase.	Glioma

그림 6. CircR2Disease 데이터 셋
Fig. 6. CircR2Disease datase

TCCTAGACAGAACCCAGGCTTCTGGGGC	hsa_circ_001621	hsa_circ_0000006
GCTTCTCGTCAGTGCCTCGCAGGATGGTA	hsa_circ_001873	hsa_circ_0000007
ATCTCACATTCTTGAAGGTGGCATTGAAG	hsa_circ_001557	hsa_circ_0000008
GTGCGGCAGCTTCCTCTTCCCCTGAGCGG	hsa_circ_000031	hsa_circ_0000009
GTGGCCAAGAAGCTGGACACACCTAATT	hsa_circ_000032	hsa_circ_0000010
GAATACTCCCCCAGTTGCAAGAGGCGCA	hsa_circ_001076	hsa_circ_0000011
AGTAAGAGGGACCATCTCCTCATGAACG	hsa_circ_000641	hsa_circ_0000012
GGCTCCAGGGAGCTTGGCTTCTGTAGAA	hsa_circ_000035	hsa_circ_0000013
GCTATTCAGGCAGCTATCAGCCAGGCCT	hsa_circ_001693	hsa_circ_0000014

그림 7. CircBase 데이터 셋
Fig. 7. CircBase dataset

에서 왼쪽의 형태를 채택하여 사용하기 때문에 본 연구에서도 왼쪽 원형 RNA symbol(hsa_circ_1234567) form를 사용하였다.

4.1.1 인간과 관련된 원형 RNA 추출

그림 8에서 Species column을 보면 인간과 관련된 데이터만 존재하는 것이 아니라 닭, 쥐와 관련된 데이터들도 존재한다. 하지만 본 연구에서는 인간과 관련된 원형 RNA와 질병의 연관성을 예측하는 연구이기 때문에 인간과 관련 없는 쥐, 닭 등의 데이터들은 삭제하였다. 삭제하는 과정에서 인간과 관련된 데이터만 남기다 보니 기존의 데이터 수에서 최대 약 700개가 줄어든 수를 확인하였다. 다른 연구에서는 다른 동물들이 포함된 데이터를 사용한다.

표 1에 표기된 데이터베이스는 원형 RNA의 Open dataset을 의미한다. 행은 각 데이터베이스에 존재하는 고유한 원형 RNA, 질병, Positive의 수(화살표 기준 왼쪽)를 의미한다. 인간과 관련된 데이터(화살표 기준 오른쪽 Bold)와 기존의 원래 각 데이터들의 수를 의미한다.

Species	circRNA name	circAltas ID	Disease Name
chicken	chr1:194735582-194744050	gal-ENSGALG00000017320_0001	Avian leukosis virus subgrou
chicken	chr5:58225735-58230830	chr5:58929908|58935003	Avian leukosis virus subgrou
chicken	chr8:1033880-1047222	gal-VAV3_0001	Avian leukosis virus subgrou
human	ARHGAP5/circARHGAP5	hsa-ARHGAP5_0003	Colon cancer
human	ARHGAP5/circARHGAP5	hsa-ARHGAP5_0003	Epithelial Ovarian carcinoma
human	CDR1as/ciRS-7/hsa_circ_0001946	hsa-CDR1_0001	Gastric cancer
human	CDR1as/ciRS-7/hsa_circ_0001946	hsa-CDR1_0001	Glioblastoma

그림 8. CircAtlas 데이터 셋
Fig. 8. CircAtlas dataaset

표 1. 기존의 Open Datasets과 인간과 관련된 데이터
Table 1. Open Datasets related to human

	MNDR 3.0[12]	CircR2Disease[9]	Circ2Disease[13]	circRNA Disease [14]	CirAtlas [11]
CircRNA	2,379 →1,742	251 →22	611 →293	330 →219	504
Disease	164 →124	60 →53	100 →58	48 →33	84
Positive	3,166 →2,389	273 →238	739 →307	354 →234	587 →581

4.1.2 통합된 CircRNA-disease 데이터베이스

본 연구에서는 CircRNA-Disease association prediction에서 주로 이용되는 데이터 셋 5가지(표 1)를 circBase[10]를 기준으로 통합하였다. 이전에 제시한 원형 RNA의 선정 문제를 circBase[10]를 통해 통일한 후 인간과 관련 없는 데이터를 삭제하여 인간과 관련된 원형 RNA와 관련된 질병 데이터셋을 구축하였다. 표 1에서 제시한 Open dataset을 통합하여 기존 연구들 보다 약 10 큰 데이터셋을 구축하였다. 표 2는 통합한 데이터셋을 정리한 표이다.

표 2에 나타난 데이터를 모델에 바로 넣어서 학습을 시킬 수 없다. 왜냐하면 입력 데이터가 벡터 형태로 표현(임베딩)되어야 하기 때문이다.

본 연구는 기존의 open dataset 명만 공개하는 기존 연구와는 다르게 각 데이터베이스에서 인간과 관련된 데이터의 수와 원형 RNA 이름을 통일 할 수 있는 기준 데이터베이스를 제시하였다.

표 2. 본 연구에서 통합된 5개의 데이터셋
Table 2. Integrated data set in this study

	Integrated dataset
Positive	2,722
CircRNA	1,886
Disease	185

4.1.3 CircRNA의 생물학적 의미를 반영하는 Embedding 기법

Embedding된 데이터들은 가상의 벡터 공간(Vector space)에 위치 시킬 수 있다. 벡터 공간에 표현된 데이터를 계산하여 알맞게 구분할 수 있는 공간 또는 결정 경계(decision boundary)를 찾는 과정을 "딥 러닝 모델이 학습한다"라고 정의한다.

그림 9는 벡터 공간에 존재하는 임베딩된 벡터(동그라미, 네모)를 나타낸 예시이다. 동그라미와 네모를 적절한 위치시키기 위해 embedding 과정이 중요하다. 왜냐하면 데이터의 특성을 반영하지 못하는 embedding 기법을 사용하면 class A에 위치할 네모 데이터가 동그라미가 위치한 class B으로 임베딩될 수 있기 때문이다. 잘못된 embedding 기법을 사용하면 모델이 복잡한 관계를 예측할 수 없게 되며, 결과적으로 모델의 성능을 떨어뜨리는 일이 된다. 그래서 입력 데이터의 특성을 알맞게 반영할 수 있는 embedding 기법[15]을 선택하는

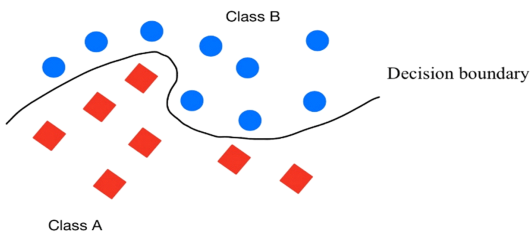

그림 9. 결정경계와 벡터공간
Fig. 9. Decision boundary in vector space

것이 중요하며, 복잡한 바이오 분야에서는 더욱 중요한 부분에 해당한다.

일반적으로 원형 RNA와 질병의 연관성을 예측하는 연구에서는 원형 RNA를 임베딩할 때 염기 서열(Sequence Base)을 주로 활용한다. 염기서열은 A,G,C,T의 조합으로 이루어져 있으며, 순서에 따라서 생물학적 의미가 달라진다. 염기서열의 배열들이 원형 RNA 특징을 나타낸다.

기존 연구에서는 임베딩된 원형 RNA가 적절한 벡터 공간에 위치하는지 즉 생물학적 의미가 알맞게 반영된 정도를 검증하지 않는다. 그래서 본 연구에서는 생물학적 의미의 반영 정도를 검증하는 부분을 제안한 Protein embedding[16] 기법을 차용하였다.

프로틴 임베딩[16]에서 다음과 같이 주장한다. 앞으로 밝혀진 모든 염기서열 정보를 계속해서 수집/정제해서 임베딩 할 수 없다. 왜냐하면 발전하는 NGS(Next-Generation Sequencing) 기술을 통해 염기서열이 끊임없이 발견되기 때문이다. 그래서 존재하는 데이터에서 생물학적인 의미를 반영하여 특징을 추출할 수 있어야 한다. 프로틴 임베딩 연구에서는 염기서열의 대표적인 특징인 localization, T50, absorption, enantioselectivity를 사용하여, 딥 러닝 모델 학습 및 생물학적 의미를 반영하는 정도를 검증하였다.

프로틴 임베딩 연구는 다음과 같은 순서로 진행된다. 그림 2에는 Unsupervised와 Supervised모델이 순차적으로 학습되는 구조로 되어있다. 그림 2에서 step 1, step 2 단계에서는 정답이 없는 염기서열 데이터에서 생물학적인 의미를 반영하여, doc2vec을 통해 embedding을 하고, 임베딩한 값을 step3에서 Test 데이터로 사용한다. 왜냐하면 Supervised 모델은 이미 염기서열의 대표적인 특징(localization, T50, absorption, enantioselectivity)을 학습하였기 때문이다.

프로틴 임베딩[16]연구은 밝혀지지 않은 염기서열에

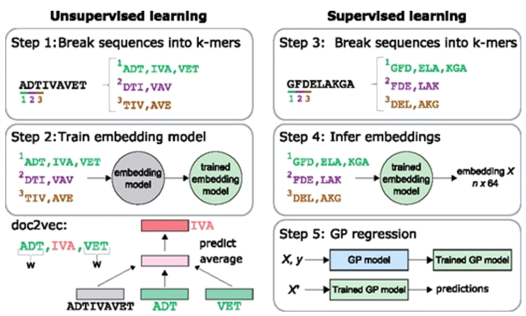

그림 10. 프로틴 임베딩 흐름
Fig. 10. Protein embedding workflow diagram

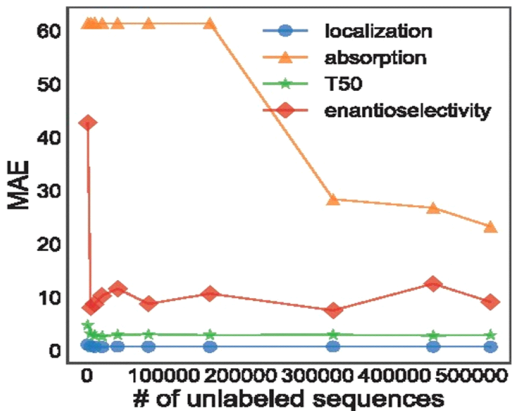

그림 11. Unlabeled 염기서열 MAE 결과
Fig. 11. MAE result of unlabeled sequence data

서도 생물학적 의미를 추출할 수 있는 능력을 가지고 있다. 그림 11에서 알 수 있듯이 염기서열의 대표적인 특징들이 MAE 값이 Unlabeled sequence가 많아질수록 오류가 줄어드는 모습을 확인할 수 있다.

그림 12에서 임베딩된 데이터의 특징을 시각화한 그림이다. 각 염기서열의 대표적인 특징들이 구분되는 모습을 확인할 수 있다.

본 연구에서 circBase[10]에 존재하는 140,790개 circRNA sequence를 염기서열의 특징을 알맞게 임베딩 할 수 있는 연구[16]의 doc2vec 모델을 사용하여, circRNA의 특징을 64차원으로 추출하였다. 그림 13은 circRNA의 특징을 추출한 결과이다.

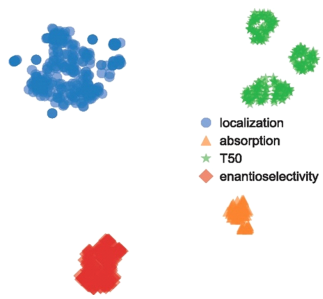

그림 12. 염기서열 특징 clustering 결과
Fig. 12. Base sequence feature clustering result

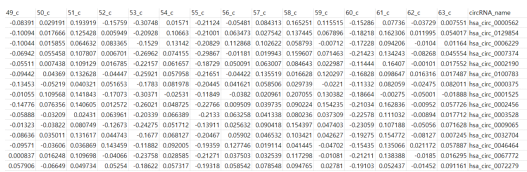

그림 13. 원형 RNA 염기서열 임베딩 결과
Fig. 13. CircRNA base sequence embedding result

4.1.4 Disease adjacency matrix 및 embedding

CircRNA-disease association 데이터 즉 Table1에 존재하는 데이터베이스에는 질병의 고유한 특징이 나타나 있지 않다. 일반적으로 질병의 특징을 나타내기 위해서 Positive 데이터를 인접 행렬(Adjacency matrix)로 나타낸다 그림 14 하나의 행은 하나의 질병을 의미하고, 하나의 열은 하나의 circRNA를 나타내며, 모든 열은 중복되지 않은 circRNA를 의미한다. 하나의 행에는 하나의 질병이므로 하나의 질병과 관련된 모든 circRNA의 관계를 알 수 있다. 앞서서 설명한 것을 그림 14에 표기하였다.

그림 14의 데이터를 GIP kernel similarity를 통해 유사도를 계산하게 된다. 다른 연구에서도 많이 활용되는 유사도 방법 중 하나이다[4,7].

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9 ..	...
d_1	1	1	0	0	1	0	1	0	1	0
d_2	0	0	1	0	0	1	0	0	0	0
⋮	0	1	0	0	0	0	1	0	1	0

그림 14. 질병의 인접 행렬
Fig. 14. Adjacency matrix of disease

V. Distance metric을 사용한 정교한 Negative set 구축 방법

5.1 기존의 랜덤으로 생성하는 Negative set 구축 방법

원형 RNA와 질병사이에 연관성이 "있다 또는 없다"를 예측하는 것은 분류 문제와 같다 그림 9에 나타난 Class A, Class B를 분류하는 문제를 이진 분류(Binary classification) 문제라 하며, 이진 분류에서 좋은 성능을 나타내는 학습 데이터(Train set)의 Positive와 Negative 비율은 1:1이다. 하지만 생물학적으로 A 와 B의 관계가 "영원히 관계가 없다" 라는 건 정의 되지 않기 때문에 Negative 데이터의 구축이 중요하다. 일반적으로 질병과 RNA연관성 예측에서 Negative set을 구축할 때 Positive 관계에서 생성될 수 있는 모든 경우의 관계에서 랜덤으로 선택하여 구축한다.

그림 15에서 원형 RNA와 관련 있는 disease 리스트를 c1-d1, c2-d2, c3-d3라고 가정하고 나타낸 예시이다. c는 원형 RNA이고, d는 disease 이다. 모든 관계 리스트에서 중복된 관계는 존재하지 않는다. 위와 같은 조건에 서 생성 될 수 있는 모든 Negative 관계는 다음과 같다. c1-d2, c1-d3, c2-d1, c2-d3, c3-d1, c3-d2. 그림

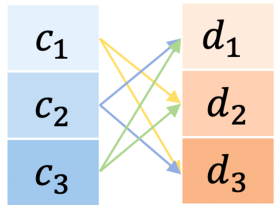

그림 15. 모든 Negative 관계 경우의 수
Fig. 15. The number of cases for all negative relationships

15는 3쌍의 데이터로 예시를 나타냈다. 하지만 본 연구에서 사용하는 실제 Train set에는 2,722(pos)쌍으로 고려해야할 Negative 관계가 기하 급수로 많다. 수 많은 후보군에서 적절한 Negative를 선택하는 방법은 좋지 못한 방법이며, 랜덤으로 선택하는 과정에서 모델이 불필요한 관계 데이터(Positive과 유사한 Negative데이터들)를 학습 및 예측하게 된다. 그래서 랜덤으로 선택하는 방법이 아닌 더욱 정교한 방법을 통해 Negative set을 구축해야 한다.

5.2 정교한 Negative set 구축 방법

일반적으로 두 데이터 간의 유사한 정도를 측정할 때 거리 함수(distance metric, distance function)를 사용한다. 본 연구에서는 총 3가지의 거리 함수 Euclidean distance(유클리드 거리, L2), Mahalanobis distance(마할라노비스 거리), Cosine similarity(코싸인 유사도)를 Positive 데이터에 적용하여 정교한 Negative set을 생성한다.

정교한 Negative set을 구축하기 위해선 먼저 Negative set에서 모델이 혼동되는 데이터를 삭제할 수 있는 기준을 지정해야 한다. 그래서 모든 Positive 관계를 거리 함수를 통해 유사도를 구한 다음 평균을 구한다. 이 평균은 Positive 데이터들의 경향을 나타며, 본 연구에서는 평균값을 기준으로 사용하였다.

그림 16은 Positive 데이터에 유클리드 거리를 적용한후 차원 축소를 통해 positive 데이터들의 분포를 확인한 모습이다. 각각의 초록색 점 하나가Positive data(circRNA-disease)이다. 그리고 빨간색 선은 유클리드 거리의 평균값을 의미한다. 그림 16의 평균값을 Negative set에 적용한 그림은 그림 17이다.

유클리드 거리는 거리가 가까울 수록 유사하다는 의미이다. 그림 17의 오른쪽에 위치한 데이터들은 확실한 Negative set이며, Positive 평균 값보다 더 짧은 거리(그림 17왼쪽)를 가지고 있는 데이터가 모델에 혼동을 주는 데이터이다. 정교한 Negative set은 이 데이터들을 삭제하였다. 랜덤으로 구축하는 Negative set은 대부분

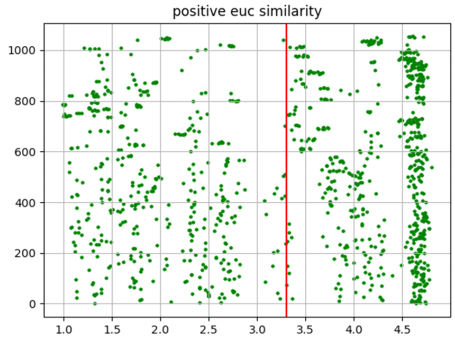

그림 16. Positive set에 Euclidean distance를 적용한 산점도 그래프
Fig. 16. Scatterplot graph with Mahalanobis distance applied to Positive set

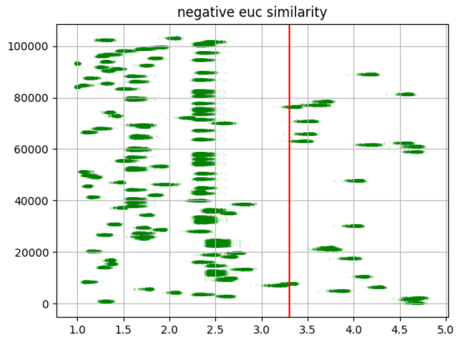

그림 17. Negative set에 Euclidean distance를 적용한 산점도 그래프
Fig. 17. Scatterplot graph with Euclidean distance applied to Negative set

평균값을 초과한 데이터를 선택하기 때문에 모델에 혼동을 주게된다.

그림 18과 그림 19는 각각 Cos similarity를 Positive 데이터와 Negative 데이터에 적용한 그림이다.

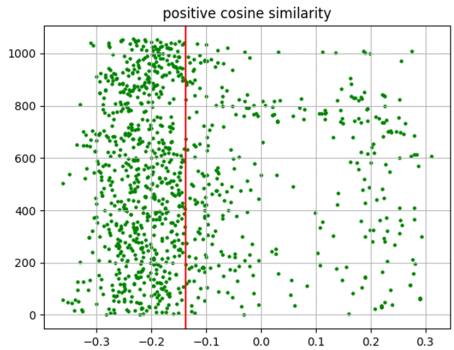

그림 18. Positive set에 Cosine similarity를 적용한 산점도 그래프
Fig. 18. Scatterplot graph with Cosine similarity applied to Positive set

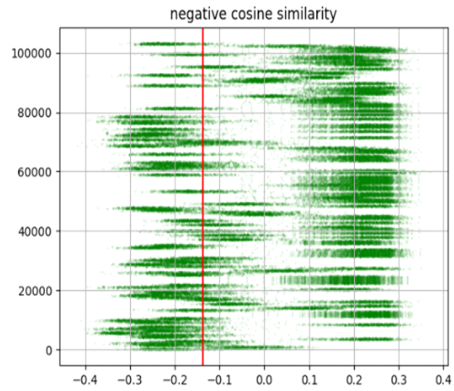

그림 19. Negative set에 Cosine similarity를 적용한 산점도 그래프
Fig. 19. Scatterplot graph with Cosine similarity applied to Negative set

Negative에서 보게되면, 수 많은 점들이 겹쳐져 있는 모습을 볼 수 있다. 코싸인 유사도는 두 벡터 사이의 각도를 계산한다. 유사할 수록 1의 값을 가지게 된다. 그림 19에서 오른쪽에 위치하고 있는 데이터가 Positive 보다 더 가까운 유사도를 가지고 있기 때문에 정교한 Negative set을 구축할 때 삭제를 하였다.

그림 20 와 그림 21은 마할라노비스 거리에 대한 결과를 나타낸다 마할라노비스는 유사할 수록 작은 값이 나오게 된다.

그림 22은 앞서 설명한 3가지 거리 함수를 통해 생성되는 Negative set의 생성 과정을 나타낸 그림이다. Embedding된 원형 RNA와 질병 벡터는 각각 64차원으로 구성되어 있다. 하나의 행(빨간 네모)이 각각의 거리 함수(유클리드 거리, 마할라노비스 거리, 코싸인 유사도)를 통해 계산하게 된다. 그 후 계산된 값을 가지고 측정한 거리함수 각각의 평균을 계산하게 된다. 평균은

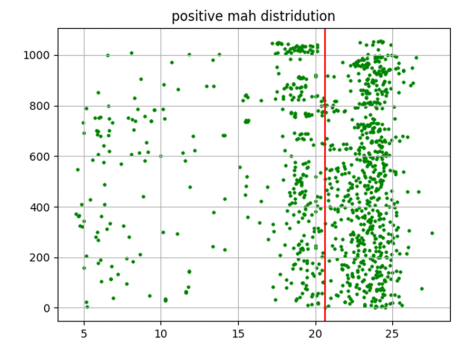

그림 20. Positive set에 Mahalanobis distance를 적용한 산점도 그래프
Fig. 20 Scatterplot graph with Mahalanobis distance applied to Positive set

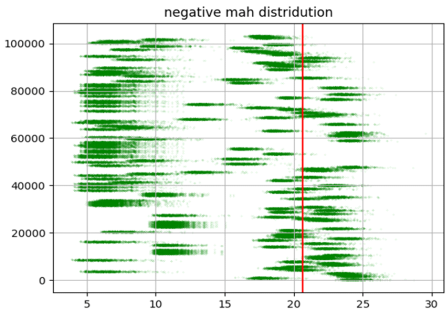

그림 21. Negative set에 Mahalanobis distance를 적용한 산점도 그래프
Fig. 21. Scatterplot graph with Mahalanobis distance applied to Negative set

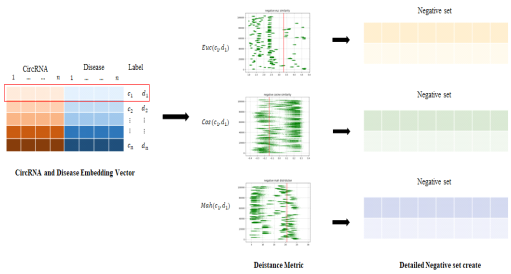

그림 22. 네거티브 셋 생성 과정
Fig. 22. Negative set generation process

Negative set의 threshold가 된다. Threshold를 기준으로 거리함수의 특징에 따라서 모델에 혼동을 줄 수 있는 negative 데이터를 제거하게 된다. 이 데이터는 앞서서 설명한 positive 와 유사한 negative를 의미한다.

Ⅵ. Detailed negative set을 사용한 DiNeg-CDA 모델 구조 및 성능 결과

본 연구에서 사용된 모델은 장단기 기억(Long Short Term Memory, LSTM) 모델이다. 순환 신경망(Recurrent Neural Network, RNN) 모델의 문제인 기울기 손실 문제(Vanishing Gradient)가 보안된 모델이다. RNN의 구조는 그림 23이다.

RNN(Recurrent Neural Network)의 기본적인 구조는 그림 23의 왼쪽에 위치한 구조가 반복되어 있다. x1x2x3…은 Sequence 데이터 즉 input data이다. x0에서 학습된 weight를 다음 x1가 들어오는 cell에 전달함으로써 이전의 정보가 새로 들어오는 정보에 반영된다. 하지만 입력 데이터가 방대해 진다면, 처음에 들어온 데이터의 파라미터가 사라지게 된다. 마치 기억력이 사라지듯이 점점 이전의 정보를 반영하지 못하게 된다. 이 현상을 기울기 소실 현상(Gradient Vanishing)이라 한다. 이 문제를 개선한 모델이 LSTM 모델이다. 본 연구에서는 LSTM 모델을 사용하였다.

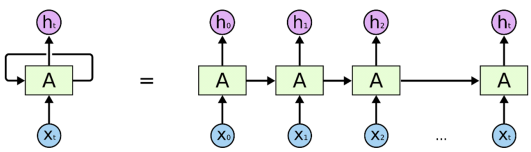

그림 23. RNN의 기본 구조
Fig. 23. Basic structure of RNN(Recurrent Neural Network)[17]

그림 24에 보게 되면 LSTM의 cell 구조이다. Forget gate에서 h_{t-1} 이전의 weight와 현재 데이터 xt가 들어오게 되고, 시그모이드 함수를 통해서 cell state에 잊어버릴 데이터가 반영된다. 그 후 input gate에 새로운 데이터가 들어오게 되고, 기억해야할 정보를 Output gate로 보내서 다음 cell로 보낼 데이터를 출력하게 된다. 앞서서 설명한 순서대로 데이터들이 셀에서 셀로 이동하면서 동작한다.

본 연구에서 그림 25의 many to many 구조로 되어

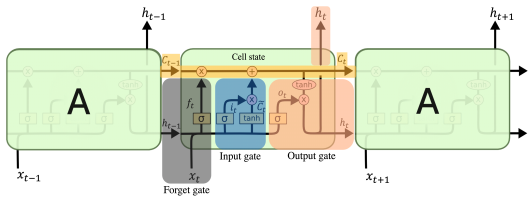

그림 24. LSTM의 기본 구조
Fig. 24. Basic of structure LSTM(Long Short Term Memory)

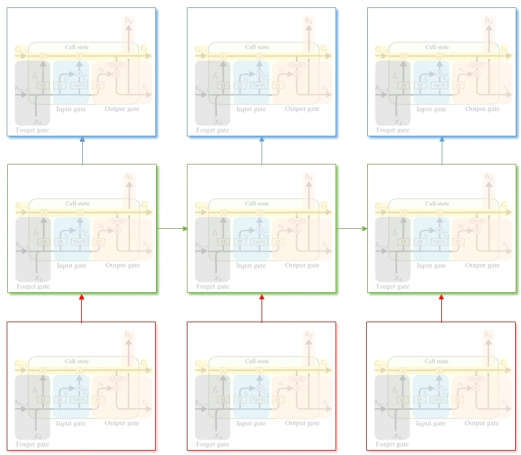

그림 25. DiNeg-CDA 모델의 many to many 구조
Fig. 25. The many-to-many structure of the DiNeg-CDA model

있으며, 하나의 관계에 대한 연관성 정보가 출력되는 구조이다. 색상이 있는 네모에는 각 lstm 모델이 며, 빨간색은 input layer 초록색은 hidden layer,파란색은 output layer이다.

그림 26은 Negative set 생성과정부터 Train 데이터가 LSTM모델에 들어가는 과정을 나타내었다.. 그림 26의 빨간 네모는 원형 RNA와 질병의 한 행은 임베딩 벡터를 의미한다. 행을 기준으로 positive와 negative으로 나뉘어져 있고, 1:1 비율로 존재한다. 임베딩 벡터들이 위에서 아래 방향으로 LSTM cell에 들어가게 된다. LSTM 모델의 레이어는 3개의 Layer로 구성 되어 있고, hidden layer는 128차원으로 되어있다. 본 모델은 Intel i7-7700 3.6Ghz, 32GB, Geforce GTX 1080Ti 2개를 달린 환경에서 실험이 이루어졌다.

모델의 성능은 Five Fold Cross Validation (FFCV)으로 성능 평가를 진행하였다. 그리고 모델의 성능 지표는 AUC ROC curve, Precision(정밀도), Recall(재현율)를 사용하였다.

정밀도는 모델이 true라고 예측한 것 중에 실제 정답(positive)인 비율을 의미한다. 재현율(recall, hit rate, sensitivity)은 정답 데이터(positive)에서 모델이 True라고 예측한 것의 비율이다. 이렇게 보게 되면 재현율과 정밀도는 유사해 보인다. 하지만 둘의 차이가 있다. 재현율은 모든 정답 데이터에서 비율을 바라보는 것이다. 그리고 정밀도는 모델의 측면에서 정답을 얼마나 맞추었는지 볼 수 있는 지표이다. 이 둘은 Trade-off 관계에 있다. 그리고 AUC ROC curve는 Roc curve에 AUC가 추가된 그래프이다. Area under cuver로 그래프의 아래의 영역이 x축 방향으로 포물선을 그리면서 그래프 아래의 영역이 넓을 수록 좋은 분류성능을 가지고 있는 모델이라고 해석 한다.

Negative set에 대한 모델의 성능을 평가 위해 3가지 거리기준을 가지고 추출한 데이터셋의 결과는 다음과 같다.

그림 27은 한 가지 거리의 기준으로 FFCV을 진행한

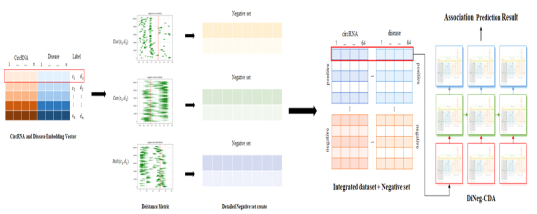

그림 26. DiNeg-CDA 모델의 many to many 구조
Fig. 26. The many-to-many structure of the DiNeg-CDA model

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	test
Euc	0.85	0.86	0.86	0.86	0.86	**0.89**
Cos	0.84	0.85	0.84	0.84	0.84	0.85
Mah	**0.89**	0.88	0.88	0.88	0.88	0.86
Random	0.85	0.85	0.85	**0.84**	0.85	**0.86**

그림 27. 한 가지 거리 기준을 사용한 Negatie set 의 5-fold 결과
Fig. 27. 5-fold result of negative set using one distance metric

결과이다. 우선 random으로 Negative를 추출하는 방법과 거리 기준으로 Negative들의 threshold를 지정하여, 삭제 했을 때가 0.89로 더 좋은 성능이 나왔다. 거리 기준으로 Negative set을 이용하여, 데이터셋을 구축하는 방법이 기존보다 CircR2Disease 의 positive 수 보다 약 10배더 커진 데이터에서도 균일한 성능 결과를 보여주었다.

그림 28과 29는 모델이 한번도 학습하지 않은 Test 데이터를 사용하여, 랜덤한 방법과 한 가지 거리기준(Single negative)을 적용한 AUC-ROC curve 결과 이다. 그림 28과 29를 비교하였을 때 유클리디안 거리를 사용한 Negative set에서 더 빠르게 학습 하는 것을 확인 할 수 있다.

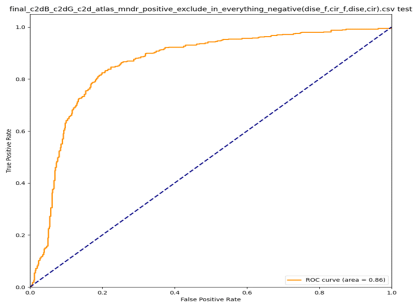

그림 28. Random Negative set의 AUC-ROC curve 결과
Fig. 28. Random Negative set AUC-ROC curve result.

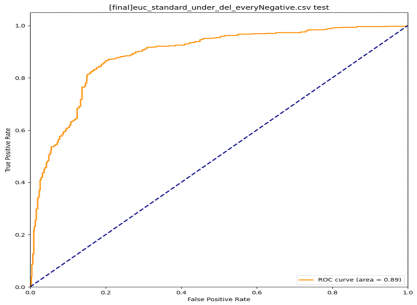

그림 29. Euclidean Negative set의 AUC-ROC curve 결과
Fig. 29. Euclidean Negative set AUC-ROC curve result

		test	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5
EUC	pre	**0.617**	0.5612	0.579	0.578	0.580	0.580
	recall	**0.917**	0.844	0.900	0.906	0.904	0.901

		test	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5
Random	pre	0.578	**0.606**	0.603	0.590	0.587	0.585
	recall	0.896	0.900	**0.907**	0.903	0.900	0.897

그림 30. Random Negative set과 EUC Negative set의 precision, recall 결과
Fig. 30. Precision, Recall result of Random Negative set and EUC Negative set

그림 30는 그림 27에서 가장 좋은 성능을 보였던 유클리드 거리를 적용한 Negative set과 랜덤으로 Negative set을 추출하는 두 방법의 정밀도와 재현율을 결과이다. Test 데이터를 사용하여, Negative를 추출하는 방법을 비교하였다. 거리기준을 적용한 모델의 재현율이 0.917로 나타났으며, 랜덤으로 추출하는 방법보다 0.021 높은 성능을 보였다. 재현율과 정밀도는 Trade-off 관계여서, 재현율이 높으면, 정밀도는 떨어진다. 거리기준을 사용한 정밀도의 결과가 더 0.617로 random을 사용한 방법보다 0.039 높은 결과를 보인다. 즉 거리기준으로 Negative set 추출 방법이 랜덤으로 추출하는 방법 보다 positive인 데이터를 positive로 예측할 수 있는 방법이다. 그리고 거리기반 Negative이 랜덤 Negative set 보다 우수한 성능을 낼 수 있는 데이터 셋이다.

더욱 정교한 Negative set을 구축하기 위해 한 가지 거리 기준(Single negative)을 거친 데이터를 다시 한번 거리기준으로 나누어 정교한 기준에 부합하는 데이터를 삭제 했을 때의 성능을 비교하였다.

그림 31은 총 두 가지의 거리기준을 차례대로 적용한 Negative set과 랜덤으로 추출한 데이터를 비교한 성능 결과이다. 5-fold의 결과에서 거리기준을 사용한 Negative set들이 랜덤으로 추출한 Negative set보다

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	test
Euc-Cos	0.88	**0.89**	0.88	0.87	0.87	**0.88**
Mah-Euc	0.87	0.85	0.85	0.85	0.85	0.87
Mah-Cos	0.87	0.86	0.85	0.86	0.86	0.86
Random	0.85	0.85	0.85	**0.84**	0.85	**0.86**

그림 31. 두 가지 Distance metric을 사용한 Negatie set과 Random Negative set의 5-fold 결과
Fig. 31. 5-fold result of Random Negtive set and using two distance metric Negative set

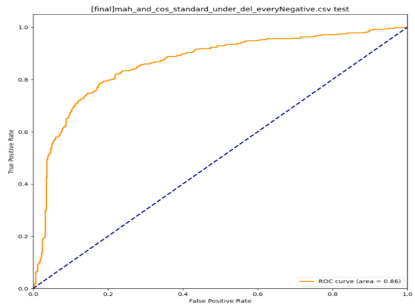

그림 32. Mah-Cos Negative set의 AUC-ROC curve
Fig. 32. AUC-ROC curve result of Mah-Cos Negative set

0.05 높은 결과를 보여줬다. 그리고 test set에서 성능이 0.02 높은 0.88을 보였다.

그림 32은 Mah-Cos를 차례 대로 적용하여 생성한 Negative set을 사용한 AUC-ROC이다.. 두 가지 거리기준과 Random negative(그림 28)를 비교하면 두가지 거리기준 negative set을 사용한 모델에서 학습 속도가 더욱 빠른 것을 보였다. 각각의 AUC는 0.86으로 동일하다.

그림 33에서 랜덤 Negative와 거리 기반 Negative와 유사한 결과를 보여준다. 하지만 전반적으로 거리 기반 Negative가 조금 더 우세한 모습이 보인다. 그리고 Test 데이터를 기준으로 볼 때 거리 기반 Negative의 재현율이 0.907로 랜덤 Negative 보다 0.011 높은 성능을 기록하였다.

두 가지 거리기준(Double negative)을 적용한 모델보다 한 가지 거리기준(Single negative)을 사용한 모델이 Test 셋에서 재현율과 정밀도 모두 조금더 우세한 성능을 보인다. 하지만 5-fold 과정에서 Double Negative가 더욱 안정적이며, 재현율 0.909로 가장 좋은 성능을 보인다. 이것은 Double Negative set이 정교한 Negative set 이다.

		test	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5
Euc-Cos	pre	**0.557**	0.5658	0.569	0.582	0.590	**0.594**
	recall	**0.907**	0.902	**0.909**	0.902	0.902	0.901

		test	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5
Random	pre	0.578	**0.606**	0.603	0.590	0.587	0.585
	recall	0.896	0.900	**0.907**	0.903	0.900	0.897

그림 33. 두 가지 Distance metric을 사용한 Negatie set과 Random Negative set의 5-fold 결과
Fig. 33. 5-fold result of Random Negtive set and using two distance metric Negative set

Ⅶ. 향후 연구 및 결론

본 연구는 심혈관 질병들과 관련이 있는 circRNA와 질병의 연관성을 예측하는 연구이다. 원형 RNA와 질병 연관성 예측 연구는 다른 RNA를 활용하여 질병을 예측하는 연구들보다 데이터가 최대 1/10 부족한 데이터 셋도 존재한다. 그래서 기존 원형 RNA와 질병 연관성 예측 연구들은 다른 데이터베이스들을 통합하여 사용한다. 하지만 각각의 데이터베이스마다 제시한 원형 RNA 및 질병 이름의 기준이 다르고, 관계(원형 RNA-질병)가 도출 된 실험과정이 다르다. 그리고 데이터베이스에는 인간과 관련된 데이터만 존재하지 않는다. 다른 동물(쥐, 박쥐, 돼지 등)들과 관련된 데이터들도 존재한다. 위에서 제시한 문제가 데이터셋을 통합하는 과정에서 문제가 된다.

본 연구에서는 앞서 말한 문제들을 고려하여 인간과 관련된 질병들만 통합한 데이터 셋 구축하였다. 구축한 데이터 셋은 기존 연구에서 활용하는 Positive 데이터 수보다 약 10배 크다.

구축한 데이터 셋(Positive set)을 가지고 딥 러닝 모델을 학습시킬 수 없기 때문에 Negative set을 생성 해야 한다. 기존 연구에서 Negative set 생성 방법은 다음과 같다. Positive set에 존재하지 않는 모든 관계 데이터(circRNA-disease)에서 랜덤으로 선택하여 학습할 Negative set을 생성한다. 하지만 이 방법은 모델의 성능을 저하시키는 요인이 된다. 랜덤으로 선택하는 방법과 본 연구에서 생성한 정밀한 Negative set을 활용하여 이를 증명하였다. Negative set(랜덤으로 선택한)을 사용한 모델은 Test set에서 recall 값이 Negative set을 사용한 모델보다 낮게 나오는 것을 확인 하였다. 이것은 Negative set의 정교한 생성과 인간과 관련된 질병을 더욱 잘 찾기 위해서 정교한 Negative set 생성(Positive 와 유사한 Negative를 삭제해 주는 작업)이 중요한 하나의 과정이다.

본 연구에서 랜덤으로 선택하는 방법과 큰 성능 차이를 보이지는 못하였다. 정교한 Negative set 생성을 위해 Threshold를 평균값을 활용하여, 진행한 부분에서 성능 저하를 불러온 것으로 예상한다. 이 부분에서 다른 cluster 모델 이나, 앙상블 학습 방법을 이용하면 좋은 성능을 보일 것이다.

중요한 정보들을 놓치지 않게 Negative set을 생성하는 연구가 필요하다. circRNA와 다른 RNA의 특징을 추출하기 위해 염기서열 데이터를 사용한다. 고유한 RNA의 특징을 추출 하는 위해 생물학적인 방법을 활용한 다양한 feature extraction 연구가 필요하다.

References

[1] D. Lee, *Understanding miRNA and LNA(2013)*, Retrieved Jul. 23, 2023, from https://www.takara.co.kr/ (http://cms.takara.co.kr/file/lsnb/1-1.pdf)

[2] M. Do, *Production and function of CircRNA (2018)*, Retrieved Jul. 23, 2023, from https://www.ibric.org/(https://www.ibric.org/myboard/read.php?Board=report&id=2919&Page=1)

[3] L.-L. Chen, “The biogenesis and emerging roles of circular RNAs,” *Nature Rev. Molecular Cell Biology*, vol. 17, no. 4, pp. 205-211, Apr. 2016. (https://doi.org/10.1038/nrm.2015.32/)

[4] S. Werfel, et al., “Characterization of circular RNAs in human, mouse and rat hearts,” *J. Molecular and Cellular Cardiology*, vol. 98, pp. 103-107, Sep. 2016. (https://doi.org/10.1016/j.yjmcc.2016.07.007)

[5] D. Jin, *The Importance of Biomarker in the Development of Anti-cancer Drugs(2016)*, Retrieved Jul. 23, 2023, from https://www.khidi.or.kr/epharmakorea(http://www.ksmcb.or.kr/board/list.html?num=4625&start=0&sort=top%20desc,pos%20asc,num%20desc&code=bbs11&period=&&title=01&key=&keyword=)

[6] D. Gwon, *Development of New Drugs Using Artificial Intelligence(2010)*, Retrieved Jul. 23, 2023, from https://www.khidi.or.kr/(https://www.khidi.or.kr/board/view?pageNum=1&rowCnt=10&no1=372&linkId=48833729&menuId=MENU01783&maxIndex=00488337629998&minIndex=00002208419998&schType=0&schText=&schStartDate=&schEndDate=&boardStyle=)

[7] C. Lu, et al., “Improving circRNA‐disease association prediction by sequence and ontology representations with convolutional and recurrent neural networks,” *Bioinformatics*, vol. 36, no. 24, pp. 5656-5664, Apr. 2021. (https://doi.org/10.1093/bioinformatics/btaa1077)

[8] L. Wang, et al., “IMS-CDA: Prediction of

CircRNA-disease associations from the integration of multisource similarity information with deep stacked autoencoder model," *IEEE Trans. Cybernetics*, vol. 51, no. 11, pp. 5522-5531, Nov. 2020. (https://doi.org/10.1109/TCYB.2020.3022852)

[9] C. Fan, et al., "CircR2Disease: A manually curated database for experimentally supported circular RNAs associated with various diseases," *The J. Biological Databases and Curation DataBase 2018*, vol. 2018, May 2018. (https://doi.org/10.1093/database/bay044)

[10] P. Glažar, P. Papavasileiou, and N. Rajewsky, "circBase: A database for circular RNAs," *Rna*, vol. 20, no. 11, pp. 1666-1670, Nov. 2014. (https://doi.org/10.1261/rna.043687.113)

[11] W. Wu, P. Ji, and F. Zhao, "CircAtlas: An integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes," *Genome biology*, vol. 21, no. 1, pp. 1-14, 2020. (https://doi.org/10.1186/s13059-020-02018-y)

[12] L. Ning, et al., "MNDR v3. 0: Mammal ncRNA–disease repository with increased coverage and annotation," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D160-D164, Jan. 2021. (https://doi.org/10.1093/nar/gkaa707)

[13] D. Yao, L. Zhang, M. Zheng, X. Sun, Y. Lu, and P. Liu, "Circ2Disease: A manually curated database of experimentally validated circRNAs in human disease," *Scientific Reports*, vol. 11018, Jul. 2018. (https://doi.org/10.1038/s41598-018-29360-3)

[14] Z. Zhao, et al., "CircRNA disease: A manually curated database of experimentally supported circRNA-disease associations," *Cell death & disease*, vol. 9, no. 5, 475, Apr. 2018. (https://doi.org/10.1038/s41419-018-0503-3)

[15] J.-H. Ha, "Autoencoder-based disease-related miRNA prediction research using deep learning," *The J. Korean Inst. Inf. Technol.*, vol. 20, no. 6, pp. 33-40, Jun. 2022. (https://doi.org/10.14801/jkiit.2022.20.6.33)

[16] K. K. Yang, et al., "Learned protein embeddings for machine learning," *Bioinformatics*, vol. 34, no. 15, pp. 2642-2648, Aug. 2018. (https://doi.org/10.1093/bioinformatics/bty178)

[17] C. Olah, *Understanding LSTM Networks cs231n(2015)*, Retrieved Jul. 23, 2023, from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

황 인 우 (In-Woo Hwang)

2021년 2월 : 목원대학교 정보통신융합공학부 졸업
2023년 2월 : 충남대학교 바이오AI융합학과 석사 졸업
<관심분야> 딥러닝, 바이오AI, 생물정보학, feature extraction, graph neural network

윤 승 원 (Seung-Won Yoon)

2018년 2월 : 충남대학교 컴퓨터공학과 졸업
2018년 3월~현재 : 충남대학교 컴퓨터공학과 석박사통합과정
<관심분야> 인공지능, 딥러닝, 머신러닝, 생물정보학

김 재 인 (Jaein Kim)

2021년 2월 : 충남대학교 컴퓨터공학과 졸업
2023년 2월 : 충남대학교 바이오AI융합학과 석사 졸업
<관심분야> 인공지능, 딥러닝, 머신러닝, 생물정보학

이 규 철 (Kyu-Chul Lee)

1984년 2월 : 서울대학교 컴퓨터 공학과 졸업
1986년 2월 : 서울대학교 컴퓨터 공학과 석사
1990년 2월 : 서울대학교 컴퓨터 공학과 박사
1989년 3월~현재 : 충남대학교 컴퓨터공학과 교수

<관심분야> 데이터베이스, 인공지능, 빅데이터
[ORCID:0000-0003-0857-807X]