

ISSN 2636-0403

발표논문집

2025
Digital
Contents
Society

(사)한국디지털콘텐츠학회
하계종합학술대회 및 대학생논문경진대회



디지털 라이프,
융합 기술과 콘텐츠

주최 | 한국디지털콘텐츠학회 Digital Contents Society 주관 | SW중심대학사업단 SKIAC MOKWON UNIVERSITY

후원 | JININFRA 위상인프라 GNSOFT 지에프티 GFT 지에프티 SJ정보통신 아이씨티웨이(주) ICTWAY 아이씨티웨이(주) CODABLE JWATS JWATS OSKO SIGONGtech (주)하이게이컨설팅 OPENLINKSYSTEM (주)에이치투테크
BonC Innovators 대진정보통신(주) NEOBRIX (주)그린텍아이앤씨 PINTEL 지엔미래교육 D PARTNERS (주)에이치투테크

사단
법인 한국디지털콘텐츠학회
Digital Contents Society

□ 논문발표 세션 WE3 (16:00~16:50)		○ 장소 : 카멜리아 홀
좌장	유선용 교수(전남대학교)	
WE3-01. 제조기업 규모와 RnD 투자에 따른 디지털 전환 격차 분석 방수경, 김진술, 이명진(전남대학교) 715		
WE3-02. YOLO v11 기술을 이용한 실시간 스킴드 탐지 모델 연구 김성권, 김진술(전남대학교) 718		
WE3-03. 연령별 안면 구분 및 맞춤형 키오스크 서비스 조혜원, 김수연, 조하영, 조영준(전남대학교) 721		
WE3-04. Transformer 기반 생물학적 그래프 모델을 활용한해석 가능한 약물 유도 유전자 발현 예측 송종웅, 서문수빈, 유선용(전남대학교) 723		
WE3-05. Line-Crossing 기반 과일봉지 계수 모델 윤경윤, 조영준(전남대학교) 725		
WE3-06. 분산신원증명을 이용한 티켓발원시스템 개선에 관한 연구 성기화(전남대학교), 안준영(㈜티맥스소프트), 김진술(전남대학교) 727		
WE3-07. ChatGPT를 활용한 디지털 콘텐츠 자동 생성의 가능성과 한계 이송배, 김진술(전남대학교) 730		
WE3-08. Conditional Diffusion Model 기반 약물로 인한 전사체 반응 예측 김채원, 정명현, 김민건, 유선용(전남대학교) 732		
WE3-09. 소규모 암시적 발화 데이터를 활용한 의도 분류 모델 성능 분석 이원주, 유연수, 김진술, 양형정(전남대학교) 734		
WE3-10. C-ITS 환경에서의 실시간 통신을 위한 REST, gRPC, RSocket 통신 프로토콜의 성능 비교 분석 및 적용 방안 연구 남현주, 옥상경, 김진술, 김경백(전남대학교) 736		
WE3-11. 디지털 지속가능성과 지속가능한 관광 간의 관계에서 인식된 경제적 이점과 친환경 인식의 조절 효과에 관한 연구 Barua Nishita, 박현재, Bithy Tahmida Amin(전남대학교) 738		

Transformer 기반 생물학적 그래프 모델을 활용한 해석 가능한 약물 유도 유전자 발현 예측

송중웅(*), 서문수빈(*), 유선용(*)

(*) 전남대학교 지능전자컴퓨터공학과, {jongwoongsong, munsu931122, syoo}@jnu.ac.kr

Interpretable Modeling of Drug-Induced Gene Expression via Biological Graphs and Transformer Encoders

Jongung Song(*), Subhin Seomun(*), Sunyong Yoo(*)

(*) Department of Intelligent Electronics and Computer Engineering

Chonnam National University, Gwangju 61186, Korea

요약

약물·세포·용량·시간을 모두 반영한 약물 유도 유전자 발현 예측은 정밀의학과 독성 평가에 필수적이다. 그러나 RNA-seq 기반 측정은 비용·시간 부담이 크고 기존 선형·기계학습 모델은 복잡한 조건 의존적 패턴을 충분히 포착하지 못한다. 본 연구는 이를 극복하기 위해 화합물 SMILES, KEGG 경로 기반 세포 그래프, 용량·시간 벡터를 Transformer 인코더로 통합한 해석 가능한 딥러닝 모델을 제안한다. 제안된 모델은 랜드마크 유전자 발현을 높은 정확도로 예측할 뿐 아니라, self-attention 메커니즘을 통해 중요한 분자 하부구조와 유전자의 기여도를 식별하고 시각화함으로써, 예측 결과의 생물학적 해석 가능성을 확보한다. 이를 통해 고비용 실험 없이도 신속한 후보 물질 탐색과 독성 평가를 가속할 것으로 기대된다.

1. 서론

유전자 발현 프로파일링은 세포가 약물 및 환경 자극에 반응하여 나타나는 전사 변화를 수천~수만 개 유전자 수준에서 고해상도로 측정하는 기술이다. 이를 활용한 Connectivity Map (Cmap)은 약물과 질병의 연관성을 밝히는 데 중요한 기여를 하고 있다 [1]. 그러나 RNA-seq는 비용과 시간이 많이 소요되어 모든 조건을 다루기 어렵고, 기존 모델은 약물 구조, 세포 특성, 용량, 시간 간의 복잡한 비선형 상호작용을 포착하거나 해석하는 데 한계가 있다 [2].

본 연구는 이를 극복하기 위해 그래프 신경망과 Transformer를 통합한 딥러닝 모델을 제안하여 다양한 조건에서 랜드마크 유전자 발현량을 정확히 예측하고, self-attention을 통해 중요한 분자 하부구조와 유전자를 시각화하여 해석 가능성을 확보한다.

2. 연구 방법

2.1 데이터 수집 및 전처리

LINCS L1000 플랫폼의 β 버전에서 약 130만 건의 시그니처를 확보하여 정제된 뒤 로그 변환 및 Z-score 정규화를 적용하였다. 추출된 데이터는 무작위로 10%를 선택한 후 80:20 비율로 학습과 검증 세트로 나눴다. 화합물은 RDKit을 이용해 SMILES를 분자 그래프로 변환했고, 세포주는 GDSC/CCLL 식별자로 정규화하였다. 용량과 시간은 정규화 후 FC Layer를 통해 임베딩 벡터로 변환하였다.

2.2 모델 설계

본 연구는 조건 의존적 유전자 발현을 정확히 예측하기 위해 그래프 신경망과 Transformer를 통합한 딥러닝 모델을 설계하였다. 화합물은 RDKit 기반의 분자

그래프로 변환한 후 GAT를 통해 원자 수준 임베딩을 생성하며, 세포주는 KEGG pathway 기반 유전자 그래프에 기저 발현값을 노드 특성으로 투영한 뒤 GATv2로 구조적 임베딩을 얻는다. 용량과 시간은 각각 32차원 벡터로 임베딩되며, 네 가지 정보는 동일한 차원으로 정규화된 뒤 Transformer 인코더에 병렬로 입력된다. Transformer의 self-attention은 서로 다른 조건 간 상호작용을 효과적으로 학습하며, 최종적으로 MLP를 통해 랜드마크 유전자의 발현량을 예측한다. 이 모델은 다양한 조건에서 높은 예측도와 함께, 경로 및 조건 단위로 해석 가능한 결과를 제공한다는 점에서 기존 모델과 차별성을 가진다.

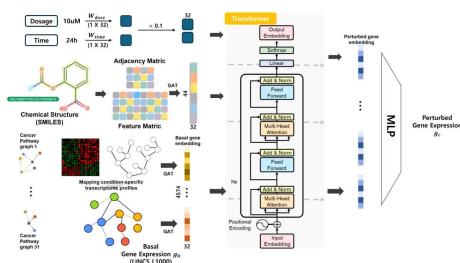


Figure 1. Proposed Model Architecture Overview

그림 1. 제안 모델 아키텍처 개요

3. 실험 결과

Linear Regression 모델은 RMSE 1.2041, R^2 에서 0.0356, PCC는 0.1493의 성능을 나타냈다. RF 모델은 RMSE 1.2034, R^2 에서 0.0370, PCC는 0.1467의 성능을 나타냈다. GAT 모델은 RMSE 1.1647, R^2 에서 0.2321, PCC는 0.2246의 성능을 나타냈다. 본 모델은 RMSE 1.1096, R^2 에서 0.3048, PCC는 0.3614의 성능을 나타냈다.

Table 1. Evaluating the performance of machine learning models

표 1. 기계학습 모델들의 성능평가

Model	RMSE	R^2	PCC
Linear (baseline)	1.2041	0.0356	0.1493
RF (baseline)	1.2034	0.0370	0.1467
GAT (baseline)	1.1647	0.2321	0.2246
Ours	1.1096	0.3048	0.3614

4. 결론 및 향후 연구

본 연구는 약물 구조, 세포 특성, 실험 조건 정보를 종합하여 약물 유도 유전자 발현을 정확히 예측할 수 있는 해석 가능한 그래프 기반 딥러닝 모델을 제안하였다. 제안 모델은 단순히 예측 정확도를 높이는 것을 넘어, 약물 작용 메커니즘 분석, 독성 평가, 가상 스크리닝 등 다양한 정밀의학 응용 가능성을 제시한다. 이번 결과는 전체 데이터 일부를 기반으로 수행한 예비 실험으로, 향후 전체 데이터셋 및 보다 다양한 실험 조건을 통합하여 모델의 일반화 성능을 강화하고, 예측 결과의 생물학적 타당성을 실험적으로 검증함으로써 모델의 신뢰성을 높일 계획이다.

감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-지역지능화혁신인재양성사업 (IITP-2025-RS-2022-00156287) 및 식품의약품안전처 연구개발비(RS-2024-00332003, RS-2025-02215961)로 수행되었으며 이에 감사드립니다.

참고 문헌 (참고자료)

- [1] Lamb, Justin et al. "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease." Science (New York, N.Y.) vol. 313,5795 (2006): 1929-35.
- [2] Subramanian, Aravind et al. "A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles." Cell vol. 171,6 (2017): 1437-1452.e17.