약-단백질 상호작용 예측과 그 불확실성 측정을 위한 뉴럴넷

제원호 교수

서울대학교 자연과학대학 물리천문학부

기술 개요

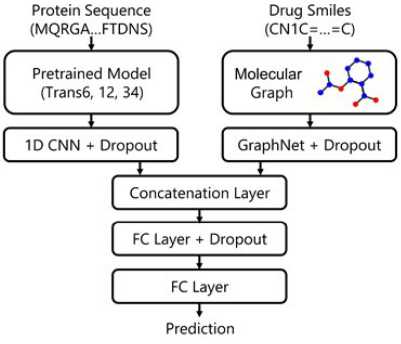
- 제약 분야에서 신약 개발을 위해 뉴럴넷 기반 인공지능을 이용하여 약물 후보 화합물과 표적 단백질 사이의 상호작용을 예측하려는 시도가 이루어지고 있음. 인공지능을 이용한 약-단백질 상호작용의 예측 성능은 약과 단백질의 표현형, 뉴럴넷 구조, 학습 방법, 학습 데이터의 품질 등에 큰 영향을 받음
- 본 기술은 약과 단백질의 특성을 잘 반영한 표현형을 이용하고 베이지안 뉴럴넷의 앙상블 학습을 통해 보다 정확한 약-단백질 상호작용을 예측하고 나아가 예측 불확실성을 측정 가능한 보다 신뢰도 높은 약-단백질 상호작용 예측 인공지능을 제안함

기술 개발 단계

- TRL4

기술 특장점

핵심기술요소	특장점
표현형 변환	• 약: 그래프 뉴럴넷을 이용하여 그래프 구조 표현형 변환 • 단백질: 거대 단백질 데이터셋으로 사전학습한 결과를 전이학습(transfer learning)한 모델을 이용하여 단백질 표현형 변환. 거대 단백질 학습 데이터를 이용해서 넓은 단백질 표현 공간 획득 가능함 • 기존 원핫 인코딩(one-hot encoding) 대비 약과 단백질의 특징을 잘 반영한 표현형 획득 가능함
드롭아웃을 이용한 베이지안 뉴럴넷	• 드롭아웃 기법을 이용하여 네트워크 앙상블을 근사적으로 구현함 • 적은 리소스(소수의 뉴럴넷)만으로 베이지안 뉴럴넷 구현 가능함
베이지안 뉴럴넷을 이용한 예측	• 베이지안 뉴럴넷을 이용하여 동시에 약-단백질 상호작용 예측 및 예측 불확실성 측정함 • 예측 불확실성으로 예측 결과의 정량적 평가 가능함 • 예측 불확실성을 고려하여 추가 데이터 확보 또는 신규 화합물 조사의 후속 조치 가능함



예측 모델 구조도

Dataset	Epistemic	Aleatoric
Human / 4	0.0128	0.020
Human / 2	0.0096	0.018
Human	0.0082	0.019
C. elegans / 4	0.0137	0.0155
C. elegans / 2	0.0098	0.0153
C. elegans	0.0053	0.0143

공공 데이터셋의 예측 불확실성 측정

Human

	Balanced Dataset (1 : 1)			Unbalanced Dataset (1 : 3)		
Methods	ROC-AUC	Precision	Recall	ROC-AUC	Precision	Recall
KNN	0.860	0.798	0.927	0.904	0.716	0.882
RF	0.940	0.861	0.897	*0.954*	0.847	0.824
L2	0.911	0.891	0.913	0.920	0.837	0.773
SVM	0.910	*0.966*	*0.950*	0.942	*0.969*	0.883
GNN	*0.970*	0.923	0.918	0.950	0.949	*0.913*
Trans6	0.968	0.902	0.901	0.971	0.915	0.910
Trans12	0.960	0.881	**0.949**	0.969	**0.958**	0.863
Trans34	0.973	0.914	0.925	0.971	0.930	0.863
Trans6+Drop	**0.975**	0.932	0.922	**0.976**	0.939	0.902
Trans12+Drop	0.971	0.914	0.924	0.963	0.932	0.902
Trans34+Drop	**0.975**	**0.945**	0.935	0.970	0.925	**0.923**

C. elegans

	Balanced Dataset (1 : 1)			Unbalanced Dataset (1 : 3)		
Methods	ROC-AUC	Precision	Recall	ROC-AUC	Precision	Recall
KNN	0.858	0.801	0.827	0.892	0.787	0.743
RF	0.902	0.821	0.844	0.926	0.836	0.705
L2	0.892	0.890	0.877	0.896	0.875	0.681
SVM	0.894	0.785	0.818	0.901	0.837	0.576
GNN	*0.978*	*0.938*	*0.929*	*0.971*	*0.916*	*0.921*
Trans6	0.981	0.937	0.949	0.977	0.871	0.917
Trans12	0.975	0.949	0.910	0.967	0.876	0.861
Trans34	0.973	0.914	0.925	0.969	0.900	0.915
Trans6+Drop	**0.986**	**0.955**	0.933	**0.983**	**0.923**	**0.944**
Trans12+Drop	0.980	0.946	0.928	0.981	0.890	0.940
Trans34+Drop	0.981	0.946	**0.940**	0.980	0.914	0.937

공공 데이터셋의 약-단백질 상호작용 예측 정확도 비교

기존 기술 현황

- 기존 인공 신경망 기반의 예측 모델
  - 인공 신경망 기능 제한됨
  - 높은 예측 모델의 복잡성
  - 비효율적인 연산 및 과적합에 취약함
  - 예측 신뢰도 산출 불가함

기존 기술대비 차별성

종래 기술의 한계	본 발명의 동작/구성	본 발명의 효과 및 이점
원핫 인코딩에 기반한 약과 단백질 표현형의 낮은 정확도	전이학습을 통한 약 표현형과 그래프 구조의 단백질 표현 이용	원핫 인코딩을 이용하지 않음에 따라 약과 단백질의 현실 구조에 가까운 표현형 변환 가능
약-단백질 상호작용 예측의 불확실성 측정 불가	베이지안 뉴럴넷을 이용한 앙상블 학습/예측	결과의 확률 분포를 앙상블하여 예측 불확실성의 측정이 가능
베이지안 뉴럴넷의 구현에 많은 리소스 필요	드롭아웃 기법을 이용하여 베이지안 뉴럴넷을 근사	소수의 뉴럴넷만으로도 베이지안 뉴럴넷을 구현할 수 있어 적은 리소스 필요

기술 활용 분야

- 제약 분야, 신물질 탐색 분야
- 인공지능 분야, 정보 분야

지식재산권 현황

No.	명칭	국가	상태	출원번호(출원일)	등록번호(등록일)	권리자
1	단백질-약 상호작용 예측과 그 불확실성 측정을 위한 베이지언 뉴럴넷 기반 인공지능	대한민국	출원	10-2021-0126415 (2021.09.24.)	-	서울대학교 산학협력단

기술 문의처

- 서울대학교 산학협력단 이한용 변리사 | 02-880-2026 | boribob@snu.ac.kr