(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년12월27일
(11) 등록번호 10-2617957
(24) 등록일자 2023년12월20일

(51) 국제특허분류(Int. Cl.)
G16C 20/30 (2019.01) *G16B 15/30* (2019.01)
G16C 20/40 (2019.01) *G16C 20/70* (2019.01)
G16C 20/80 (2019.01)
(52) CPC특허분류
G16C 20/30 (2019.02)
G16B 15/30 (2019.02)
(21) 출원번호 10-2023-0049561
(22) 출원일자 2023년04월14일
심사청구일자 2023년04월14일
(56) 선행기술조사문헌
M. Li et al., BACPI: a bi-directional attention neural network for compound-protein interaction and binding affinity prediction, Bioinformatics Vol.38, pp1995-2002(2022)*
Q. Zhao et al., HyperAttentionDTI: improving drug-protein interaction prediction by sequence-based deep learning with attention mechanism, Bioinformatics Vol.38, pp655-662(2021)*
S. Li et al., MONN: A Multi-objective Neural Network for Predicting Compound-Protein Interactions and Affinities, Cell Systems Vol.10, pp308-322, e1-e11(2020)*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
고려대학교산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
강재우
서울특별시 성동구 왕십리로 241 (행당동, 서울숲더샵) 103동 1402호
김모건
서울특별시 서초구 사임당로17길 116 (서초동 서초삼성래미안아파트) 103동 1302호
(뒷면에 계속)
(74) 대리인
특허법인위솔

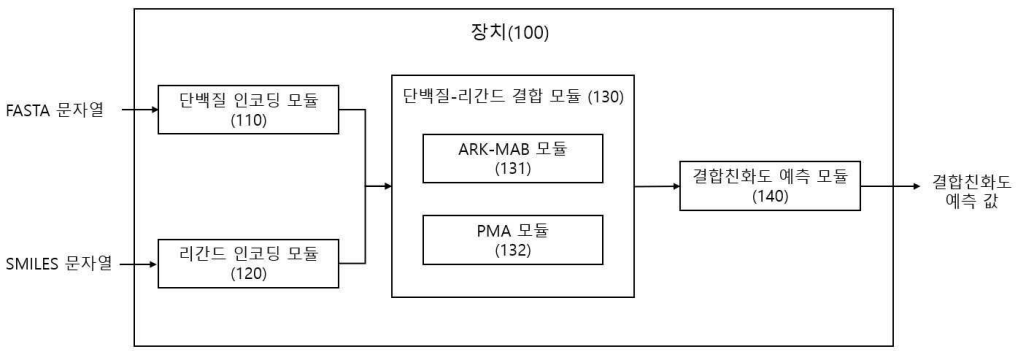
전체 청구항 수 : 총 8 항 심사관 : 권계민

(54) 발명의 명칭 **비공유결합 기반 단백질-리간드 결합 친화도 예측 방법 및 장치**

(57) 요 약

본 발명의 일 실시 예에 따른 비공유결합 기반 단백질-리간드 결합친화도 예측 장치는, 입력된 단백질을 기반으로 m 개의 아미노산 잔기 단위의 표현형 벡터를 생성하는 단백질 인코딩 모듈, 입력된 리간드를 기반으로 n 개의 부분구조 단위의 표현형 벡터를 생성하는 리간드 인코딩 모듈, 상기 m 개의 아미노산 잔기 단위의 표현형 벡터 및 상기 n 개의 부분구조 단위의 표현형 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 백터를 생성하는 단백질-리간드 결합 모듈 및 상기 단백질-리간드 결합체에 대한 단일 표현형 벡터를 기반으로 상기 단백질과 상기 리간드의 결합친화도를 예측하는 결합친화도 예측 모듈을 포함할 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G16C 20/40 (2019.02)
G16C 20/70 (2019.02)
G16C 20/80 (2019.02)

(72) 발명자
최준석
서울특별시 동대문구 무학로34가길 3 404호
백승흔
서울시 동대문구 무학로 삼영오피스텔 101호
박주언
서울특별시 서초구 효령로 68길 81 서초자이아파트 101동 1203호
이채은
경기도 의정부시 평화로150번길 26, 204동 703호
주민재
성북구 고려대로 13길 26 반석빌딩 504호
이수민
서울특별시 노원구 동일로 986 101-1101

이 발명을 지원한 국가연구개발사업
과제고유번호 1711193663
과제번호 2020-0-01819-004
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 ICT명품인재양성(고려대학교)
기 여 율 20/100
과제수행기관명 고려대학교산학협력단
연구기간 2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711188732
과제번호 2023R1A2C3004176
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 개인기초연구(과기정통부)
연구과제명 생명의료 가설 Co-Design을 위한 Human-in-the-loop 초대형 언어모델 개발
기 여 율 40/100
과제수행기관명 고려대학교
연구기간 2023.03.01 ~ 2024.02.29
이 발명을 지원한 국가연구개발사업
과제고유번호 1465039951
과제번호 HR20C0021030023
부처명 보건복지부
과제관리(전문)기관명 한국보건산업진흥원
연구사업명 연구중심병원육성
연구과제명 인공지능·빅데이터 기반 데이터주도 의료기술 개발 및 고도화
기 여 율 40/100
과제수행기관명 고려대학교산학협력단
연구기간 2023.01.01 ~ 2023.12.31

명 세 서

청구범위

청구항 1

비공유결합 기반 단백질-리간드 결합친화도 예측 장치에 있어서,

입력된 단백질을 기반으로 m 개의 아미노산 잔기 단위의 표현형 벡터를 생성하는 단백질 인코딩 모듈;

입력된 리간드를 기반으로 n 개의 부분구조 단위의 표현형 벡터를 생성하는 리간드 인코딩 모듈;

상기 m 개의 아미노산 잔기 단위의 표현형 벡터 및 상기 n 개의 부분구조 단위의 표현형 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 벡터를 생성하는 단백질-리간드 결합 모듈; 및

상기 단백질-리간드 결합체에 대한 단일 표현형 벡터를 기반으로 상기 단백질과 상기 리간드의 결합친화도를 예측하는 결합친화도 예측 모듈;

을 포함하는 장치에 있어서,

상기 단백질-리간드 결합 모듈은,

상기 m 개의 아미노산 잔기 단위의 표현형 벡터, 상기 n 개의 부분구조 단위의 표현형 벡터 및 학습 가능한 하나의 가상 부분구조 표현형 벡터를 기반으로 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터를 생성하는 ARK-MAB(Attention Regularization based on non-covalent interactions in multihead attention block) 모듈; 및

상기 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터 및 시드 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 벡터를 생성하는 PMA(Pooling by multihead attention layer) 모듈을 포함하며,

상기 ARK-MAB 모듈은,

상기 m 개의 아미노산 잔기 단위의 표현형 벡터를 기반으로 쿼리 프로젝션(query projection)을 생성하고,

상기 n 개의 부분구조 단위의 표현형 벡터 및 상기 하나의 가상 부분구조 표현형 벡터를 기반으로 키 프로젝션(key projection) 및 밸류 프로젝션(value projection)을 생성하는 것을 특징으로 하는 장치.

청구항 2

삭제

청구항 3

삭제

청구항 4

제1항에 있어서,

상기 AKR-MAB 모듈은,

상기 쿼리 프로젝션 및 상기 키 프로젝션을 기반으로 어텐션 가중치(attention weight) 행렬을 생성하고,

상기 어텐션 가중치 행렬의 각 행의 합이 1이 되도록 소프트맥스 정규화를 적용하고,

상기 소프트맥스 정규화가 적용된 어텐션 가중치 행렬 및 밸류 프로젝션을 기반으로 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터를 생성하는 것을 특징으로 하는 장치.

청구항 5

제1항에 있어서,

상기 ARK-MAB 모듈은 각 아미노산 잔기와 리간드 사이에 화학적 비공유결합이 존재하는지 여부를 나타내도록 미리 정의된 비공유결합(non-covalent interaction, NCI) 레이블을 기반으로 학습되는 것을 특징으로 하는 장치.

청구항 6

제5항에 있어서,

상기 NCI 레이블은,

특정 아미노산 잔기와 상기 리간드 사이에 화학적 비공유결합이 존재할 경우 1의 값을 가지고,

상기 특정 아미노산 잔기와 상기 리간드 사이에 화학적 비공유결합이 존재하지 않을 경우 0의 값을 가지는 것을 특징으로 하는 장치.

청구항 7

제6항에 있어서,

제1 아미노산 잔기에 대응되는 NCI 레이블의 값이 1일 경우, 상기 ARK-MAB 모듈은 상기 제1 아미노산 잔기에 대해 상기 n 개의 부분구조 단위의 표현형 벡터에 대응되는 어텐션 가중치의 합이 높아지고 상기 하나의 가상 부분구조 표현형 벡터에 대응되는 어텐션 가중치가 낮아지도록 학습되는 것을 특징으로 하는 장치.

청구항 8

제6항에 있어서,

제2 아미노산 잔기에 대응되는 NCI 레이블의 값이 0일 경우, 상기 ARK-MAB 모듈은 상기 제2 아미노산 잔기에 대해 상기 하나의 가상 부분구조 표현형 벡터에 대응되는 어텐션 가중치가 높아지고 상기 n 개의 부분구조 단위의 표현형 벡터에 대응되는 어텐션 가중치의 합이 낮아지도록 학습되는 것을 특징으로 하는 장치.

청구항 9

비공유결합 기반 단백질-리간드 결합친화도 예측 방법에 있어서,

(a) 단백질 인코딩 모듈이 입력된 단백질을 기반으로 m 개의 아미노산 잔기 단위의 표현형 벡터를 생성하는 단계;

(b) 리간드 인코딩 모듈이 입력된 리간드를 기반으로 n 개의 부분구조 단위의 표현형 벡터를 생성하는 단계;

(c) 단백질-리간드 결합 모듈이 상기 m 개의 아미노산 잔기 단위의 표현형 벡터 및 상기 n 개의 부분구조 단위의 표현형 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 벡터를 생성하는 단계; 및

(d) 결합친화도 예측 모듈이 상기 단백질-리간드 결합체에 대한 단일 표현형 벡터를 기반으로 상기 단백질과 상기 리간드의 결합친화도를 예측하는 단계;

를 포함하는 방법에 있어서,

상기 (c) 단계는,

(c-1) 상기 m 개의 아미노산 잔기 단위의 표현형 벡터, 상기 n 개의 부분구조 단위의 표현형 벡터 및 학습 가능한 하나의 가상 부분구조 표현형 벡터를 기반으로 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터를 생성하는 단계; 및

(c-2) 상기 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터 및 시드 벡터를 기반으로 단백질-리간드 결합체에

대한 단일 표현형 벡터를 생성하는 단계;

를 포함하며,

(c-3) 상기 m 개의 아미노산 잔기 단위의 표현형 벡터를 기반으로 쿼리 프로젝션(query projection)을 생성하는 단계; 및,

(c-4) 상기 n 개의 부분구조 단위의 표현형 벡터 및 상기 하나의 가상 부분구조 표현형 벡터를 기반으로 키 프로젝션(key projection) 및 밸류 프로젝션(value projection)을 생성하는 단계;

를 더 포함하는 것을 특징으로 하는 방법.

청구항 10

컴퓨팅 장치와 결합하여,

(A) 입력된 단백질을 기반으로 m 개의 아미노산 잔기 단위의 표현형 벡터를 생성하는 단계;

(B) 입력된 리간드를 기반으로 n 개의 부분구조 단위의 표현형 벡터를 생성하는 단계;

(C) 상기 m 개의 아미노산 잔기 단위의 표현형 벡터 및 상기 n 개의 부분구조 단위의 표현형 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 벡터를 생성하는 단계; 및

(D) 상기 단백질-리간드 결합체에 대한 단일 표현형 벡터를 기반으로 상기 단백질과 상기 리간드의 결합친화도를 예측하는 단계;

를 실행시키기 위하여 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램에 있어서,

상기 (C) 단계는,

(C-1) 상기 m 개의 아미노산 잔기 단위의 표현형 벡터, 상기 n 개의 부분구조 단위의 표현형 벡터 및 학습 가능한 하나의 가상 부분구조 표현형 벡터를 기반으로 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터를 생성하는 단계; 및

(C-2) 상기 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터 및 시드 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 벡터를 생성하는 단계;

를 포함하며,

(C-3) 상기 m 개의 아미노산 잔기 단위의 표현형 벡터를 기반으로 쿼리 프로젝션(query projection)을 생성하는 단계; 및,

(C-4) 상기 n 개의 부분구조 단위의 표현형 벡터 및 상기 하나의 가상 부분구조 표현형 벡터를 기반으로 키 프로젝션(key projection) 및 밸류 프로젝션(value projection)을 생성하는 단계;

를 더 포함하는 것을 특징으로 하는 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램.

발명의 설명

기 술 분 야

[0001] 본 발명은 비공유결합 기반 단백질-리간드 결합 친화도 예측 방법 및 장치에 관한 것이며, 보다 상세하게는 신약 개발을 위해 설명가능한 단백질-리간드 결합 친화도 예측력을 갖춘 인공지능 모델을 활용한 비공유결합 기반의 단백질-리간드 결합 친화도 예측 방법 및 이를 위한 장치에 관한 것이다.

배 경 기 술

[0003] 약물-표적 상호작용(Drug-target interaction, DTI)의 식별은 약물 설계 및 개발의 핵심 과제 중 하나이다. 기존의 생체 내(in-vivo) 및 생체 외(in-vitro) 실험에 기초한 전통적인 약물 개발 프로세스는 많은 비용과 노동이 소모되기 때문에, 현재는 딥러닝을 통하여 단백질-리간드 간의 결합 친화도를 예측하는 연구들이 진행되고 있다.

[0004] 그러나 제한된 딥러닝 모델의 설명 가능성은 이러한 예측모델을 실제 산업에 적용하는데 장애물로 남아있다. 최근에는 이와 같은 설명 가능성 문제를 해결하기 위해 어텐션-기반(attention-based) 방법이 적용되고 있다. 이와 함께 모델에게 중요한 생물 및 화학적 지식을 학습시켜 전문가의 의견과 서로 양립할 수 있는 방향으로 연구가 진행되고 있다. 그러나 기존의 단백질-리간드 결합친화도 예측 모델은 학습 데이터로부터 사용자에게 유용하거나 설명 가능한 어텐션-기반 방법을 갖추기 어렵다. 이는 단백질-리간드 결합친화도 데이터의 부족 및 실제 예측의 난이도에서 비롯된 문제다.

[0005] 한편, DTI 연구에서 단백질-리간드 사이의 화학적 비공유결합(non-covalent interaction, NCI) 개념은 모델이 이해해야 하는 대표적인 개념이다. 비공유결합은 단백질과 리간드가 어떻게 서로 상호작용하고 복합체를 형성하는지 이해하는데 필수적이며, 이는 약물 화합물의 작용 메커니즘에 영향을 미친다. 대부분의 약물 화합물은 리간드로 작용하고 단백질과 같은 생물학적 표적과 비공유결합을 형성함으로써 치료효과를 나타낸다.

[0006] 본 발명은 단백질-리간드 사이의 화학적 비공유결합을 고려한 모델 설명력을 갖춘 단백질-리간드 결합친화도 예측 방법에 대한 것이다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 대한민국 등록특허공보 제10-2237189호(2021.04.01.)

발명의 내용

해결하려는 과제

[0009] 본 발명이 해결하고자 하는 기술적 과제는 단백질-리간드 사이의 화학적 비공유결합을 고려한 모델 설명력을 갖춘 단백질-리간드 결합친화도 예측 방법을 제공하는 것이다.

[0010] 본 발명의 기술적 과제들은 이상에서 언급한 기술적 과제들로 제한되지 않으며, 언급되지 않은 또 다른 기술적 과제들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0012] 전술한 기술적 과제를 달성하기 위한 본 발명의 일 실시 예에 따른 비공유결합 기반 단백질-리간드 결합친화도 예측 장치는, 입력된 단백질을 기반으로 m 개의 아미노산 잔기 단위의 표현형 벡터를 생성하는 단백질 인코딩 모듈, 입력된 리간드를 기반으로 n 개의 부분구조 단위의 표현형 벡터를 생성하는 리간드 인코딩 모듈, 상기 m 개의 아미노산 잔기 단위의 표현형 벡터 및 상기 n 개의 부분구조 단위의 표현형 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 백터를 생성하는 단백질-리간드 결합 모듈 및 상기 단백질-리간드 결합체에 대한 단일 표현형 벡터를 기반으로 상기 단백질과 상기 리간드의 결합친화도를 예측하는 결합친화도 예측 모듈을 포함할 수 있다.

[0013] 일 실시 예에 따르면, 상기 단백질-리간드 결합 모듈은, 상기 m 개의 아미노산 잔기 단위의 표현형 벡터, 상기 n 개의 부분구조 단위의 표현형 벡터 및 학습 가능한 하나의 가상 부분구조 표현형 벡터를 기반으로 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터를 생성하는 ARK-MAB(Attention Regularization based on non-covalent interactions in multihead attention block) 모듈 및 상기 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터 및 시드 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 벡터를 생성하는 PMA(Pooling by multihead attention layer) 모듈을 포함할 수 있다.

[0014] 일 실시 예에 따르면, 상기 ARK-MAB 모듈은, 상기 m 개의 아미노산 잔기 단위의 표현형 벡터를 기반으로 쿼리 프로젝션(query projection) 생성하고, 상기 n 개의 부분구조 단위의 표현형 벡터 및 상기 하나의 가상 부분구조 표현형 벡터를 기반으로 키 프로젝션(key projection) 및 밸류 프로젝션(value projection)을 생성할 수 있다.

[0015] 일 실시 예에 따르면, 상기 AKR-MAB 모듈은, 상기 쿼리 프로젝션 및 상기 키 프로젝션을 기반으로 어텐션 가중치(attention weight) 행렬을 생성하고, 상기 어텐션 가중치 행렬의 각 행의 합이 1이 되도록 소프트맥스 정규화를 적용하고, 상기 소프트맥스 정규화가 적용된 어텐션 가중치 행렬 및 밸류 프로젝션을 기반으로 m 개의 정

제된 아미노산 잔기 단위의 표현형 벡터를 생성할 수 있다.

[0016] 일 실시 예에 따르면, 상기 ARK-MAB 모듈은 각 아미노산 잔기와 리간드 사이에 화학적 비공유결합이 존재하는지 여부를 나타내도록 미리 정의된 비공유결합(non-covalent interaction, NCI) 레이블을 기반으로 학습될 수 있다.

[0017] 일 실시 예에 따르면, 상기 NCI 레이블은, 특정 아미노산 잔기와 상기 리간드 사이에 화학적 비공유결합이 존재할 경우 1의 값을 가지고, 상기 특정 아미노산 잔기와 상기 리간드 사이에 화학적 비공유결합이 존재하지 않을 경우 0의 값을 가질 수 있다.

[0018] 일 실시 예에 따르면, 제1 아미노산 잔기에 대응되는 NCI 레이블의 값이 1일 경우, 상기 ARK-MAB 모듈은 상기 제1 아미노산 잔기에 대해 상기 n 개의 부분구조 단위의 표현형 벡터에 대응되는 어텐션 가중치의 합이 높아지고 상기 하나의 가상 부분구조 표현형 벡터에 대응되는 어텐션 가중치가 낮아지도록 학습될 수 있다.

[0019] 일 실시 예에 따르면, 제2 아미노산 잔기에 대응되는 NCI 레이블의 값이 0일 경우, 상기 ARK-MAB 모듈은 상기 제2 아미노산 잔기에 대해 상기 하나의 가상 부분구조 표현형 벡터에 대응되는 어텐션 가중치가 높아지고 상기 n 개의 부분구조 단위의 표현형 벡터에 대응되는 어텐션 가중치의 합이 낮아지도록 학습될 수 있다.

[0020] 또한, 본 발명의 일 실시 예에 따른 비공유결합 기반 단백질-리간드 결합친화도 예측 방법은 단백질 인코딩 모듈이 입력된 단백질을 기반으로 m 개의 아미노산 잔기 단위의 표현형 벡터를 생성하는 단계, 리간드 인코딩 모듈이 입력된 리간드를 기반으로 n 개의 부분구조 단위의 표현형 벡터를 생성하는 단계, 단백질-리간드 결합 모듈이 상기 m 개의 아미노산 잔기 단위의 표현형 벡터 및 상기 n 개의 부분구조 단위의 표현형 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 백터를 생성하는 단계 및 결합친화도 예측 모듈이 상기 단백질-리간드 결합체에 대한 단일 표현형 벡터를 기반으로 상기 단백질과 상기 리간드의 결합친화도를 예측하는 단계를 포함할 수 있다.

[0021] 또한, 본 발명의 일 실시 예에 따른 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램은 입력된 단백질을 기반으로 m 개의 아미노산 잔기 단위의 표현형 벡터를 생성하는 단계, 입력된 리간드를 기반으로 n 개의 부분구조 단위의 표현형 벡터를 생성하는 단계, 상기 m 개의 아미노산 잔기 단위의 표현형 벡터 및 상기 n 개의 부분구조 단위의 표현형 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 백터를 생성하는 단계 및 상기 단백질-리간드 결합체에 대한 단일 표현형 벡터를 기반으로 상기 단백질과 상기 리간드의 결합친화도를 예측하는 단계를 포함할 수 있다.

발명의 효과

[0023] 본 발명의 다양한 실시 예들에 따르면, 주어진 미지의 표적 단백질에 대해 어떤 아미노산 잔기 부분이 실질적으로 리간드와 화학적 비공유결합을 형성하는지를 사용자에게 제공할 수 있다.

[0024] 또한, 본 발명의 다양한 실시 예들에 따르면, 해당 미지의 표적 단백질의 결합 부위(binding pocket)을 사용자에게 간접적으로 제공할 수 있다.

[0025] 또한, 본 발명의 다양한 실시 예들에 따르면, 약물 화합물인 리간드에 대해 어떤 화학적 부분구조가 표적 단백질에 대해 더 상대적으로 많은 화학적 영향력을 행사하고 있는 지를 사용자에게 제공할 수 있다.

[0026] 또한, 본 발명의 다양한 실시 예들에 따르면, 약물 화합물 설계에 있어서 중요한 부분을 알려줌으로써 신약 개발 파이프라인에 기여할 수 있는 효과가 있다.

[0027] 본 발명의 효과들은 이상에서 언급한 효과들로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0029] 도 1은 본 발명의 일 실시 예에 따른 비공유결합 기반 단백질-리간드 결합 친화도 예측을 위한 장치의 구성을 나타낸 도면이다.

도 2는 본 발명의 일 실시 예에 따른 ARK-MAB 모듈의 구성을 나타낸 도면이다.

도 3은 본 발명의 일 실시 예에 따른 헤드 별 어텐션 연산 수행 과정을 타나낸 도면이다.

도 4는 하나의 헤드에서 어텐션 연산이 수행되는 과정을 도식화한 도면이다.

도 5는 복수의 헤드에 대해 수행된 어텐션 연산의 결과를 결합하는 것을 도시한 도면이다.

도 6은 본 발명의 일 실시 예에 따른 PMA 모듈의 구성을 나타낸 도면이다.

도 7은 PMA 모듈에서 생성된 군집화된 표현형 벡터에 선형 연산을 적용하여 단일 단백질-리간드 결합체 표현형 벡터를 생성하는 과정을 나타낸 도면이다.

도 8은 본 발명의 일 실시 예에 따른 NCI 기반의 보조 손실함수를 이용한 학습 방법을 나타내기 위한 도면이다.

도 9는 다양한 모델들을 이용한 결합친화도 예측의 성능을 나타낸 도면이다.

도 10은 본 발명의 일 실시 예에 따른 비공유결합 기반 단백질-리간드 결합 친화도 예측 방법에 기반한 어텐션 맵의 일 예시를 나타낸 도면이다.

도 11은 본 발명의 일 실시 예에 따른 비공유결합 기반 단백질-리간드 결합 친화도 예측 방법에 기반한 어텐션 맵의 다른 일 예시를 나타낸 도면이다.

도 12는 본 발명의 일 실시 예에 따른 비공유결합 기반 단백질-리간드 결합 친화도 예측 방법에 기반한 어텐션 맵의 또 다른 일 예시를 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0030] 본 발명의 목적과 기술적 구성 및 그에 따른 작용 효과에 관한 자세한 사항은 본 발명의 명세서에 첨부된 도면에 의거한 이하의 상세한 설명에 의해 보다 명확하게 이해될 것이다. 첨부된 도면을 참조하여 본 발명에 따른 실시 예를 상세하게 설명한다.

[0031] 본 명세서에서 개시되는 실시 예들은 본 발명의 범위를 한정하는 것으로 해석되거나 이용되지 않아야 할 것이다. 이 분야의 통상의 기술자에게 본 명세서의 실시 예를 포함한 설명은 다양한 응용을 갖는다는 것이 당연하다. 따라서, 본 발명의 상세한 설명에 기재된 임의의 실시 예들은 본 발명을 보다 잘 설명하기 위한 예시적인 것이며 본 발명의 범위가 실시 예들로 한정되는 것을 의도하지 않는다.

[0032] 도면에 표시되고 아래에 설명되는 기능 블록들은 가능한 구현의 예들일 뿐이다. 다른 구현들에서는 상세한 설명의 사상 및 범위를 벗어나지 않는 범위에서 다른 기능 블록들이 사용될 수 있다. 또한, 본 발명의 하나 이상의 기능 블록이 개별 블록들로 표시되지만, 본 발명의 기능 블록들 중 하나 이상은 동일 기능을 실행하는 다양한 하드웨어 및 소프트웨어 구성들의 조합일 수 있다.

[0033] 또한, 어떤 구성요소들을 포함한다는 표현은 "개방형"의 표현으로서 해당 구성요소들이 존재하는 것을 단순히 지칭할 뿐이며, 추가적인 구성요소들을 배제하는 것으로 이해되어서는 안 된다.

[0034] 나아가 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급될 때에는, 그 다른 구성요소에 직접적으로 연결 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 한다.

[0035] 도 1은 본 발명의 일 실시 예에 따른 비공유결합 기반 단백질-리간드 결합 친화도 예측을 위한 장치의 구성을 나타낸 도면이다. 이하 본 발명을 설명함에 있어서 "장치"(100)는 특별히 정의하지 않는 한 도 1을 참조하여 설명한 비공유결합 기반 단백질-리간드 결합 친화도 예측을 위한 장치를 의미하는 것으로 이해될 수 있다. 다만 도 1의 예시는 본 발명의 목적을 달성하기 위한 바람직한 실시 예일 뿐이며, 필요에 따라 일부 구성이 추가되거나 삭제될 수 있고, 어느 한 구성이 수행하는 역할을 다른 구성이 함께 수행할 수도 있음은 물론이다.

[0036] 도 1을 참조하면, 본 발명의 일 실시 예에 따른 장치(100)는 단백질 인코딩 모듈(110), 리간드 인코딩 모듈(120), 단백질-리간드 결합 모듈(130) 및 결합 친화도 예측 모듈(140)을 포함할 수 있다.

[0037] 단백질 인코딩 모듈(110)은 입력 표적 단백질을 아미노산 잔기 단위의 표현형 벡터로 변환할 수 있다. 표적 단백질은 아미노산 잔기 서열의 특성 상 20 가지 알파벳의 조합으로 이루어진 문자열로 표현될 수 있는데, 이처럼 표적 단백질의 아미노산 서열을 알파벳의 조합을 통해 나타낸 문자열을 FASTA 문자열이라고 한다. FASTA 문자열은 일반적으로 500자 내지 1000자 사이의 길이를 가지며, 문자열의 각 알파벳은 특정 아미노산 잔기를 나타낸다. 이러한 FASTA 문자열은 표적 단백질의 아미노산 서열을 나타내는 원시적인 표현형으로서, 이를 모델에서 인식 가능하고 행렬 기반의 연산이 가능한 형태로 변환할 필요가 있다. 단백질 인코딩 모듈(110)은 표적 단백질의 원시적인 표현형인 FASTA 문자열을 입력받아 이를 아미노산 잔기 단위의 표현형 벡터로 변환하여 출력하는 기능을 수행한다. 이를 위한 단백질 인코딩 모듈(110)은 단백질의 구조 및 생화학적 특성을 예측하며, 억

단위의 단백질 구조 데이터를 학습한 모델로서 다양한 단백질 표현형 벡터 및 그에 속한 아미노산 잔기 단위의 표현형 벡터들을 생성하는 범용성을 가지는 ESM2(evolutionary scale modeling 2) 모델을 기반으로 구현될 수 있으나, 이는 본 발명의 범위를 한정하지 않음에 유의한다.

[0038] 상기와 같이, 단백질 인코딩 모듈(110)은 m개의 아미노산 잔기를 가지는 표적 단백질에 대한 원시적인 표현형으로서 길이 m인 아미노산 서열(FASTA 문자열)을 입력 받으며, 해당 입력에 대한 출력으로서 차원 길이 d를가지는 아미노산 잔기 단위의 m 개의 표현형 벡터들의 집합 R을 출력할 수 있다. 여기서 출력 R은 (m X d) 크기의 행렬로 표현될 수 있다.

[0039] 리간드 인코딩 모듈(120)는 입력 리간드를 화학적 부분구조 단위의 표현형 벡터로 변환할 수 있다. 리간드는 약물 화합물의 분자 구조의 특성을 나타내는 SMILES라는 문자열로 표현될 수 있는데, 이러한 SMILES 문자열은 리간드를 나타내는 원시적인 표현형으로서 이를 모델에서 인식 가능하고 행렬 기반의 연산이 가능한 형태로 변환할 필요가 있다. 리간드 인코딩 모듈(120)은 이러한 SMILES 문자열을 입력받아 이를 화학적 부분구조 단위의 표현형 단백질을 출력하는 기능을 수행한다.

[0040] 리간드 인코딩 모듈(120)는 우선 RDKit을 활용하여 입력 데이터인 SMILES 문자열을 이진 벡터인 Morgan Fingerprint로 변환한다. 여기서 Morgan Fingerprint는 1024로 고정된 길이를 가지며, 0 또는 1로 이루어진 이진 벡터이다. Morgan Fingerprint의 각 자리의 0 또는 1의 값은 입력 리간드에 특정 화학적 부분구조가 존재하는지 여부를 나타낼 수 있다. Morgan Fingerprint에서 "1"은 해당 "1"의 값을 가지는 자리에 대응되는 특정 화학적 부분구조가 입력 리간드에 존재함을 나타내고, "0"은 해당 "0"의 값을 가지는 자리에 대응되는 특정 화학적 부분구조가 입력 리간드에 존재하지 않음을 나타낸다. 예를 들어, 만약 Morgan Fingerprint에서 13, 53, 593, 1021 번째 자리의 값이 1인 경우, 입력 리간드는 13, 53, 593, 1021 번째 자리에 상응하는 화학적 부분구조로 이루어져 있다고 간주될 수 있다.

[0041] 본 발명은 Morgan Fingerprint의 길이와 동일한 1024 개의 학습 가능한 표현형 벡터들의 집합인 색인 테이블(Lookup table)을 정의한다. 해당 테이블의 각 표현형 벡터들은 차원 길이 d를 가지며, 색인 테이블을 구성하는 각 표현형 벡터들은 특정 화학적 부분구조를 나타내도록 학습 가능한 데이터로 설계된다. 즉 색인 테이블은 화학적 부분구조를 나타내는 1024 개의 학습 가능한 표현형 벡터들의 집합으로 정의된다.

[0042] 리간드 인코딩 모듈(120)은, 색인 테이블에서 Morgan Fingerprint의 값이 1에 해당하는 표현형 벡터들을 추출하며 추출된 표현형 벡터들의 집합을 출력할 수 있다. 예를 들어, Morgan Fingerprint에서 13, 53, 593, 1021 번째 자리의 값이 1인 경우, 리간드 인코딩 모듈(120)은 색인 테이블에서 13, 53, 593, 1021 번째 표현형 벡터들을 추출하여 출력한다.

[0043] 상기와 같이, 리간드 인코딩 모듈(120)은 리간드를 나타내는 SMILES 문자열을 입력받으며, 해당 입력에 대한 출력으로 미리 정의된 색인 테이블로부터 추출된 n개의 표현형 벡터들의 집합 S를 출력할 수 있다. 여기서 출력 S는 (n X d) 크기의 행렬로 표현될 수 있으며, 이 때 n은 입력받은 SMILES 문자열로부터 생성된 Morgan Fingerprint에서 1의 값을 가지는 자리의 수이면서 동시에 입력 리간드가 가지는 화학적 부분 구조의 개수를 의미할 수 있다.

[0044] 단백질-리간드 결합 모듈(130)은 단백질 인코딩 모듈(110)을 통해 생성된 아미노산 잔기 단위의 표현형 벡터들의 집합 R 및 리간드 인코딩 모듈(120)을 통해 생성된 화학적 부분구조 단위의 표현형 벡터들의 집합 S를 입력받아, 단백질-리간드 결합체에 대한 단일 표현형 백터를 생성하여 출력할 수 있다. 단백질-리간드 결합 모듈(130)은 두 개의 서브 모듈인 ARK-MAB(Attention Regularization based on non-covalent interactions in multihead attention block) 모듈(131) 및 PMA(Pooling by multihead attention layer) 모듈(132)을 포함할 수 있다.

[0045] 도 2는 본 발명의 일 실시 예에 따른 ARK-MAB 모듈의 구성을 나타낸 도면이다.

[0046] ARK-MAB 모듈(131)은 단백질 인코딩 모듈(110) 및 리간드 인코딩 모듈(120)으로부터 m 개의 아미노산 잔기 단위의 표현형 벡터 집합 R과 n 개의 화학적 부분구조 단위의 표현형 벡터 집합 S를 입력받아 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터를 생성하여 출력한다. ARK-MAB 모듈(131)은 기존 트랜스포머(transformer)에 적용되는 교차 어텐션(cross-attention) 기반 방법론을 화학적 비공유결합(non-covalent interactions, NCI)이 고려되도록 개선한 모듈로서, MAB(multihead attention block) 및 RFF(row-wise feedforward layer)로 구성될 수 있다.

[0047] ARK-MAB 모듈(131)의 MAB는 복수 개의 헤드(head) 별로 어텐션 연산을 수행하며, 각 헤드 별 연산 결과들을 서

로 결합하여 최종 결과를 출력한다. 이하 본 발명에서는 예시로 헤드의 개수가 4개인 것을 예시로 설명할 것이나, 이는 본 발명의 범위를 제한하지 않음에 유의한다.

[0048] 도 3은 본 발명의 일 실시 예에 따른 헤드 별 어텐션 연산 수행 과정을 타나낸 도면이고, 도 4는 하나의 헤드에서 어텐션 연산이 수행되는 과정을 도식화한 도면이다.

[0049] 310 단계에서, ARK-MAB 모듈(131)의 MAB는 리간드 인코딩 모듈(120)로부터 입력된 화학적 부분구조 단위의 표현형 벡터 집합 S에, 학습 가능하도록 정의된 가상 부분구조 표현형 벡터 p를 추가하여 새로운 집합 S'을 생성할 수 있다. 이 때 S가 (n X d)의 크기를 가지는 행렬로 표현되므로, 하나의 가상 부분구조 표현형 벡터 p가 추가된 S'은 (n+1 X d)의 크기를 가질 수 있다. 본 발명에서 S에 추가되는 가상 부분구조 표현형 벡터 p는 ARK-MAB 모듈(131)이 각 아미노산 잔기 중 실질적으로 리간드와 화학적 비공유결합을 형성하는 아미노산 잔기와 화학적 비공유결합을 형성하지 않는 아미노산 잔기를 구분할 수 있도록 보조 손실함수를 학습시키기 위해 사용되는 가상의 벡터이며, 가상 부분구조 표현형 벡터 p를 활용한 손실함수의 학습 방법은 아래에서 후술하도록 한다.

[0050] 320 단계에서, ARK-MAB 모듈(131)의 MAB는 단백질 인코딩 모듈(110)로부터 입력된 표현형 벡터 집합 R 및 앞서 410 단계에서 생성된 S'을 기반으로 쿼리 프로젝션(query projection) Q, 키 프로젝션(key projection) K, 밸류 프로젝션(value projection) V를 생성할 수 있다. 여기서 쿼리 프로젝션 Q는 R을 기반으로 생성되며, (d X d/4) 크기를 가지는 행렬 W_q를 기반으로 Q = R X W_q와 같이 생성될 수 있다. R은 (m X d) 크기의 행렬로 나타낼 수 있으므로, Q는 (m X d/4) 크기의 행렬일 수 있다. 키 프로젝션 K는 (d X d/4) 크기를 가지는 행렬 W_k를 기반으로 K = S'X W_k와 같이 생성될 수 있다. 밸류 프로젝션 V는 (d X d/4) 크기를 가지는 행렬 W_v를 기반으로 V = S'X W_v와 같이 생성될 수 있다. S'는 (n+1 X d)의 크기를 가지는 행렬로 나타낼 수 있으므로, K 및 V는 각각 (n+1 X d/4)의 크기를 가지는 행렬로 나타낼 수 있다. 320 단계에서 생성된 Q, K, V는 각각 표현형 벡터의 차원 길이 d를 헤드의 개수 4로 나눈 d/4의 차원 길이를 가지도록 생성된다.

[0051] 330 단계에서, ARK-MAB 모듈(131)의 MAB는 쿼리 프로젝션 Q를 구성하는 아미노산 잔기 단위의 표현형 벡터와 키 프로젝션 K를 구성하는 화학적 부분구조 단위의 표현형 벡터의 상대적인 관계성을 수치화한 값인 어텐션 가중치(attention weight) 행렬 W를 생성할 수 있다. 이를 위해 MAB는 합산 어텐션(additive Attention)Q와 K를 구성하는 각 행 벡터들 간의 Scoring 연산을 수행할 수 있다. 결과로 생성된 어텐션 가중치 행렬 W는 (m X n+1)의 크기를 가지며, W의 각 행은 특정 아미노산 잔기와 n+1 개의 화학적 부분구조와의 상대적인 관계성을 나타낸다.

[0052] 340 단계에서, ARK-MAB 모듈(131)의 MAB는 (m X n+1) 크기를 가지는 어텐션 가중치 행렬 W의 각 행에 있는 값들의 합이 1이 되도록 소프트맥스 정규화(softmax normalization)을 적용하여 W와 동일한 (m X n+1) 크기를 가지는 행렬 W'를 생성할 수 있다.

[0053] 350 단계에서, ARK-MAB 모듈(131)의 MAB는 소프트맥스 정규화가 적용된 행렬 W'에 밸류 프로젝션 V를 곱하여 n+1개의 스칼라곱이 적용된 value 표현형 벡터들에 대해 가중 합(weighted sum)을 적용하여 Q'을 산출하면 하나의 헤드에 대한 어텐션 연산이 종료된다. 여기서 V 행렬은 (n+1 X d/4)의 크기를 가지므로, 하나의 헤드에 대해 수행된 어텐션 연산을 통해 산출되는 Q' 행렬은 (m X d/4)의 크기를 가진다.

[0054] 도 5는 복수의 헤드에 대해 수행된 어텐션 연산의 결과를 결합하는 것을 도시한 도면이다.

[0055] 도 3 및 도 4의 예시를 참조하면, 헤드 별 어텐션 연산 과정은 4개의 헤드 Head#1 내지 Head #4에 대해서 각각 수행되며, 결과로 4개의 헤드에 대해 각각 (m X d/4) 크기의 행렬 Q' 4개가 생성된다. 4개의 헤드에 대해 생성된 4 개의 (m X d/4) 크기의 행렬을 도 5에 도시된 바와 같이 횡방향으로 부착시키면 (m X d) 크기의 행렬이 O를 생성된다. ARK-MAB 모듈(131)의 MAB는 차원 길이가 d인 m개의 표현형 벡터를 나타내는 행렬 O를 최종 출력값으로 출력한다.

[0056] 이상 도 3 내지 도 5를 통해 ARK-MAB 모듈(131)의 MAB를 이용한 어텐션 연산 과정을 설명하였다. 다시 도 2를 참조하면, ARK-MAB 모듈(131)은 ARK-MAB 모듈(131)의 MAB의 최종 출력인 O과 표적 단백질의 아미노산 잔기 단위의 표현형 벡터 집합 R에 대해 합산 연산을 수행한 후, 레이어 정규화(Layer Normalization, LayerNorm)를 적용하여 R'을 생성한다. 마지막으로 ARK-MAB 모듈(131)은 상기 R'을 RFF(row-wise feedforward layer)에 통과시켜 생성된 R'_rff와 R'에 대해 합산 연산을 수행한 후 다시 레이어 정규화(Learner)을 적용하여 m개의 정제된 아미노산 잔기 단위의 표현형 벡터들의 집합인 R"을 최종 출력으로 출력하며, ARK-MAB 모듈(131)의 최종 출력 R"은 PMA 모듈(132)의 입력 데이터로 사용된다.

[0057] PMA 모듈은(132)은 ARK-MAB 모듈(131)에서 생성되는 m개의 정제된 아미노산 잔기 단위의 표현형 벡터 집합 R"을

입력받으며, 하나의 차원 길이가 d인 결합체 표현형 벡터를 생성하여 출력한다.

[0058] 도 6은 본 발명의 일 실시 예에 따른 PMA 모듈의 구성을 나타낸 도면이다. 도 6을 참조하면, PMA 모듈(132)는 ARK-MAB 모듈(131)과 유사하게 MAB(multihead attention block) 및 RFF(row-wise feedforward layer)로 구성될 수 있다. PMA 모듈(132)에서는 2개의 학습 가능한 차원 길이가 d인 시드(seed) 벡터 U가 탑재되어 있으며, 이를 기반으로 쿼리 프로젝션 Q를 생성한다. 또한 PMA 모듈(132)에서 키 프로젝션 K는 ARK-MAB 모듈(131)로부터 입력받은 m개의 정제된 아미노산 잔기 단위의 표현형 벡터 R"를 기반으로 키 프로젝션 K 및 밸류 프로젝션 V를 생성할 수 있다. 이후 쿼리 프로젝션 Q, 키 프로젝션 K 및 밸류 프로덕션 V를 이용한 어텐션 연산은 앞서 설명한 ARK-MAB 모듈(131)과 동일하며, 여기서는 상세한 설명은 생략한다. PMA 모듈(132)은 시드 벡터의 수에 해당하는 차원 길이 d인 2개의 군집화된 표현형 벡터를 생성하며, 생성된 2개의 군집화된 표현형 벡터는 도 7에 도시된 바와 같이 횡방향으로 부착시킨 후, 차원 축소를 목적으로 한 선형(linear) 연산을 적용하여 차원 길이가 d인 단일 단백질-리간드 결합체 표현형 벡터로 변환된다. 해당 단일 단백질-리간드 결합체 표현형 벡터가 최종 출력 값으로 출력되며, 결합친화도 예측 모듈(140)에 입력되어 결합친화도 값을 예측하기 위해 사용된다.

[0059] 결합친화도 예측 모듈(140)은 단백질-리간드 결합 모듈(130)의 PMA 모듈(131)에서 출력된 차원 길이가 d인 단일 단백질-리간드 결합체 표현형 벡터를 기반으로 최종적으로 예측되는 결합친화도 값을 생성하여 출력할 수 있다.

[0060] 본 발명은 ARK-MAB 모듈(131)의 각 아미노산 잔기 중 실질적으로 약물 화합물인 리간드와 화학적 비공유결합(NCI)을 형성하는 것과 형성하지 않는 것을 구분할 수 있도록 고도화하기 위해, 가상 표현형 벡터 p를 포함하여 생성된 어텐션 가중치 행렬을 이용하는 보조 손실함수를 도입하였다. 보조 손실함수는 특정 아미노산 잔기가 단백질-리간드 결합체를 이룰 때, 해당 리간드와의 비공유결합을 이루지 않을 경우 화학적 부분구조들에 해당 특정 아미노산 잔기와의 어텐션 가중치(attention weight)를 분배할 때 실질적인 부분구조가 아닌 가상의 표현형 벡터 p에 할당되도록 모델을 학습시키고, 반대로 특정 아미노산 잔기가 단백질-리간드 결합체를 이룰 때 해당 리간드와의 비공유결합을 이룰 경우에는 어텐션 가중치가 가상의 표현형 벡터 p가 아닌 실질적인 부분구조들에 할당되도록 모델을 학습시킨다.

[0061] 도 8은 본 발명의 일 실시 예에 따른 NCI 기반의 보조 손실함수를 이용한 학습 방법을 나타내기 위한 도면이다.

[0062] 앞서 ARK-MAB 모듈(131)을 설명하면서, m개의 아미노산 잔기 단위의 표현형 벡터 집합 R로부터 생성된 쿼리 프로젝션 및 n개의 화학적 부분구조 단위의 표현형 벡터 집합 S에 가상 표현형 벡터 p를 추가한 집합 S'으로부터 생성된 키 프로젝션을 기반으로 (m X n+1) 크기의 어텐션 가중치(attention weight) 행렬이 생성되고, 이후 소프트맥스 정규화가 적용될 수 있음을 설명하였다. 이하 도 8을 참조한 설명에서 어텐션 가중치 행렬은 각 행의 값들의 합이 1이 되도록 소프트맥스 정규화가 적용된 행렬을 의미하는 것으로 이해될 수 있다.

[0063] 도 8에서 $q_1, \cdots, q_i$은 ARK-MAB 모듈(131)에서 생성된 어텐션 가중치 행렬의 각 행(row)의 인덱스를 나타내며, 각각 m개의 표적 단백질의 아미노산 잔기로부터 생성된 쿼리(query)에 대응된다. 또한, 도 8에서 $k_1, \cdots, k_n$ 및 k_p는 어텐션 가중치 행렬의 각 열(column)의 인덱스를 나타내며 여기서 $k_1, \cdots, k_n$는 각각 입력 리간드의 화학적 부분구조로부터 생성된 키(key)에 대응되고, k_p는 가상 부분구조 표현형 벡터 p에 대응된다.

[0064] 도 8을 참조하면, 본 발명은 어텐션 가중치 행렬에서 각 q_i(i=0,1,$\cdots$,m) 행 별로 k_1 열 내지 k_n 열에 존재하는 어텐션 값들을 모두 합산하여 새로운 $\sum k$ 열을 정의한다. 이 때 합산은 실제 입력 리간드의 화학적 부분구조에 대응되는 k_1 열 내지 k_n 열의 어텐션 값들에 대해서만 수행되며, 가상 부분구조 표현형 벡터에 대응되는 k_p 열의 값은 합산되지 않음에 유의한다. 앞서 설명한 바와 같이 도 8의 어텐션 가중치 행렬은 소프트맥스 정규화가 적용된 상태이므로, 각 행의 값들의 합은 1이며 따라서 각 행 별로 $\sum k$가 1에 가까울수록 k_p는 0에 가까운 값을 가지고, $\sum k$가 0에 가까울수록 k_p는 1에 가까운 값을 가진다.

[0065] 합산된 $\sum k$ 값은 각 q_i 행에 대응되는 아미노산 잔기와 입력 리간드 간 화학적 비공유결합 존재 여부에 대한 예측 값으로 간주되며, 예를 들어 q_i 행의 $\sum k$ 열 값이 1에 가까울수록 해당 q_i 행에 대응되는 아미노산 잔기와

입력 리간드 간 화학적 비공유결합이 존재할 가능성이 높고, $\sum k$ 열 값이 0에 가까울수록 해당 q_i에 대응되는 아미노산 잔기와 입력 리간드 간 화학적 비공유결합이 존재하지 않을 가능성이 높은 것으로 간주될 수 있다.

[0066] 따라서, 본 발명은 특정 아미노산 잔기와 입력 리간드가 화학적 비공유결합을 형성할 경우 ARK-MAB 모듈(131)이 $\sum k$ 값이 1에 가까워지도록 어텐션 가중치 행렬을 생성하고, 특정 아미노산 잔기와 입력 리간드가 화학적 비공유결합을 형성하지 않을 경우 ARK-MAB 모듈(131)이 $\sum k$ 값이 0에 가까워지도록 어텐션 가중치 행렬을 생성하도록 학습되는 보조 손실함수를 제안한다.

[0067] 상기 보조 손실함수를 이용한 학습을 위해, 본 발명은 각 아미노산 잔기와 입력 리간드 사이에 실제 화학적 비공유결합이 존재하는지 여부를 나타내는 NCI 레이블 정보를 정의한다. NCI 레이블 정보의 각 NCI 레이블은 각 아미노산 잔기와 리간드 사이에 화학적 비공유결합이 존재하는지 여부에 따라 0 또는 1의 값을 가지도록 정의되며, 0은 해당 아미노산 잔기와 리간드 사이에 화학적 비공유결합이 존재하지 않음을 의미하고, 1은 해당 아미노산 잔기와 리간드 사이에 화학적 비공유결합이 존재함을 의미하며 따라서 NCI 레이블은 보조 손실함수를 이용한 학습을 위한 ground-truth 값으로 활용될 수 있다.

[0068] 예를 들어, 도 8의 예시에서는 NCI 레이블 값이 0인 q_1, q_2, q_m에 대응되는 아미노산 잔기는 입력 리간드와 화학적 비공유결합이 존재하지 않음을 나타내고, NCI 레이블 값이 1인 q_3에 대응되는 아미노산 잔기는 입력 리간드와 화학적 비공유결합이 존재함을 나타낸다.

[0069] 본 발명에서 제안하는 보조 손실함수는 상기 NCI 레이블을 기반으로 각 아미노산 잔기가 입력 리간드와 화학적 비공유 결합이 존재하지 않을 경우 어텐션 가중치를 분배할 때 실제 리간드의 화학적 부분구조에 대응되는 $\sum k$가 아닌 가상의 부분구조 p에 대응되는 k_p에 높은 어텐션 가중치가 분배되도록 모델을 학습시킨다. 반대로, 본 발명에서 제안하는 보조 손실함수는 상기 NCI 레이블을 기반으로 각 아미노산 잔기가 입력 리간드와 화학적 비공유 결합이 존재할 경우에는 가상의 부분구조 p에 대응되는 k_p가 아닌 실제 리간드의 화학적 부분구조에 대응되는 $\sum k$에 높은 어텐션 가중치가 분배되도록 모델을 학습시킨다.

[0070] 이상 도 8을 참조하여 설명한 보조 손실함수는 결합친화도 예측 모듈(140)을 통해 출력되는 최종 결합친화도 예측 값과 실제 결합친화도 값의 차이를 최소화하는 메인(main) 손실함수 외에 연산 과정에서 ARK-MAB 모듈(131)에 의해 생성되는 어텐션 가중치 행렬을 이용하는 보조적인 손실함수로 사용되며, 본 발명은 이처럼 두 개의 손실함수를 통해 결합 친화도를 더욱 정확하게 예측할 수 있는 방법을 제공한다. 또한, 보조 손실함수는 ARK-MAB 모듈(131)에 의해 생성되는 어텐션 가중치가 실제 화학적 비공유결합이 존재하는지 여부를 나타내는 NCI 레이블 정보 학습에 사용되므로 기존 기술과 비교할 때 화학적으로 모델 설명력을 갖도록 할 수 있다.

[0071] 도 9는 다양한 모델들을 이용한 결합친화도 예측의 성능을 나타낸 도면이다.

[0072] 도 9에서 ArkDTA가 본원 발명의 일 실시 예에 따른 비공유결합 기반 단백질-리간드 결합 친화도 예측 방법을 이용한 방법의 성능을 나타내며, 그 외 DeepDTA,GraphDTA, TransformerCPI, MONN, HyperAttentionDTI, AttentionDTA, BACPI, HFDTI 등 여러 모델에 대해 성능을 나타내고 있다. 성능을 측정하기 위한 데이터로는 PDBbind 데이터셋에서 측정 단위에 따른 다양한 샘플들 중 KIKD 타입의 샘플과 IC50 타입의 샘플을 사용하였으며, 다시 측정 방식에 따라 PDBbind KIKD, PDBbind KIKD (Unseen Scaffolds), PDBbind IC50, PDBbind IC50 (Unseen Scaffolds) 네 종류의 데이터에 대하여 여러 모델들의 성능을 나타낸다. 도 9에서 볼 수 있듯이, ArkDTA는 성능을 나타내는 지표들 RMSE(root mean squared error), MAE(mean absolute error), PCORR(partial correlation) 및 CI(C index)에 대해서 여러 모델 들과 비교할 때 좋은 성능을 가짐을 확인할 수 있다.

[0073] 도 10 내지 도 12는 본 발명의 일 실시 예에 따른 비공유결합 기반 단백질-리간드 결합 친화도 예측 방법에 기반한 어텐션 맵을 나타낸 도면이다.

[0074] 도 10 내지 도 12와 같이, 어텐션 맵을 통해 사용자에게 특정 단백질의 어떤 부분이 화학적 결합에 기여하고 있고, 특정 분자의 어떠한 부분 구조가 더 많은 영향력을 행사하고 있는지를 직관적으로 확인할 수 있도록 한다.

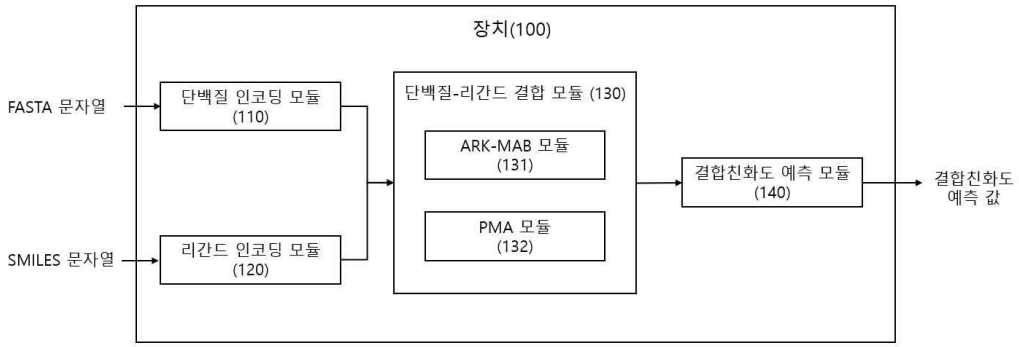
[0075] 본 발명의 일 실시 예에 따른 비공유결합 기반 단백질-리간드 결합 친화도 예측 방법은 동일한 기술적 특징을 포함하는 매체에 저장된 컴퓨터 프로그램으로 구현할 수 있는바, 이 경우 컴퓨팅 장치와 결합하여,입력된 단백질을 기반으로 m 개의 아미노산 잔기 단위의 표현형 벡터를 생성하는 단계, 입력된 리간드를 기반으로 n 개의

부분구조 단위의 표현형 벡터를 생성하는 단계, 상기 m 개의 아미노산 잔기 단위의 표현형 벡터 및 상기 n 개의 부분구조 단위의 표현형 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 백터를 생성하는 단계 및 상기 단백질-리간드 결합체에 대한 단일 표현형 벡터를 기반으로 상기 단백질과 상기 리간드의 결합친화도를 예측하는 단계를 포함할 수 있으며, 중복 서술을 위해 자세히 설명하지는 않았지만 본 발명에 적용된 모든 기술적 특징은 컴퓨터 프로그램에 모두 동일하게 적용될 수 있다.

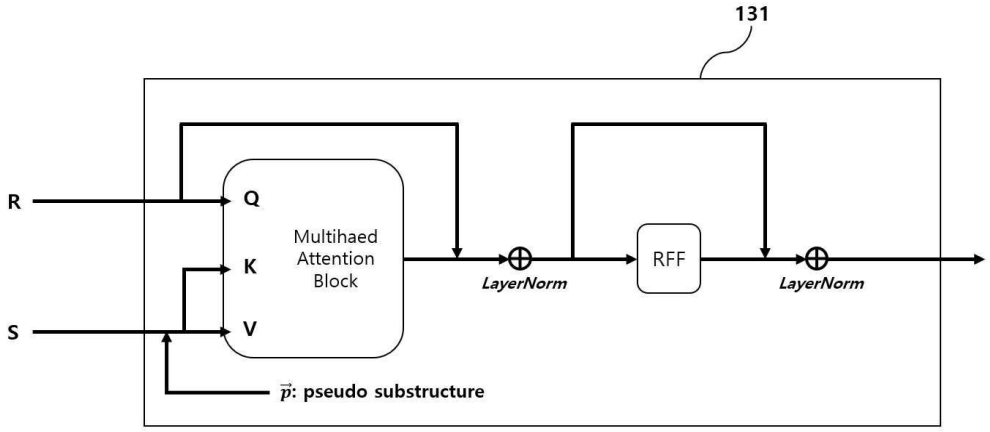
[0076] 이상 첨부된 도면을 참조하여 본 발명의 실시 예들을 설명하였지만, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시 예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다.

도면

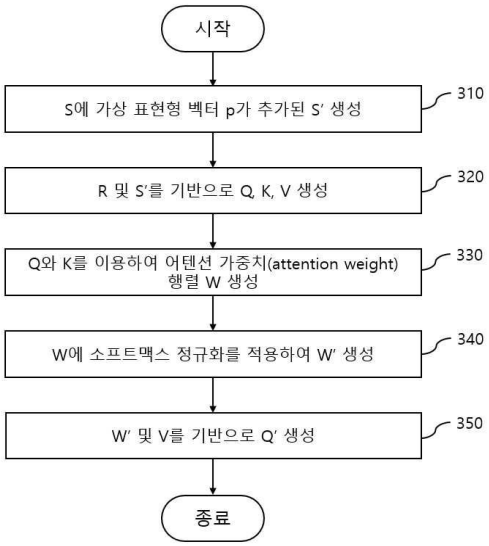
도면1



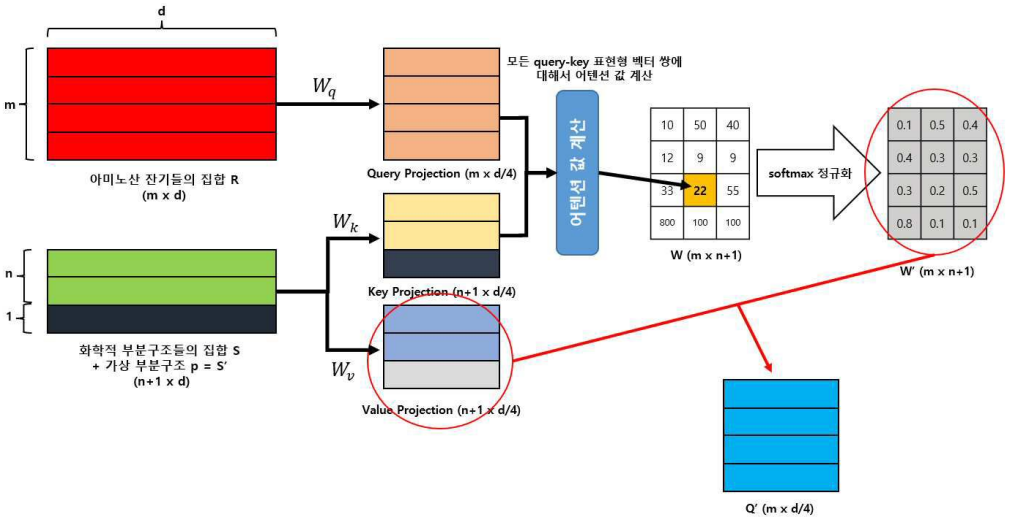
도면2



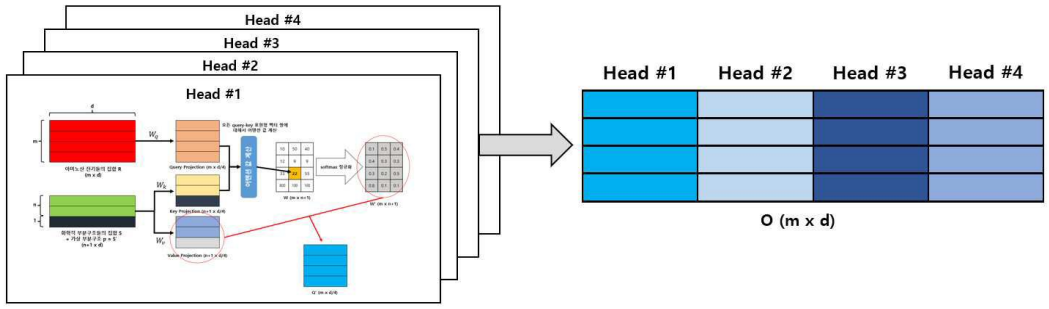
도면3


시작
S에 가상 표현형 벡터 p가 추가된 S′ 생성
310
R 및 S′를 기반으로 Q, K, V 생성
320
Q와 K를 이용하여 어텐션 가중치(attention weight) 행렬 W 생성
330
W에 소프트맥스 정규화를 적용하여 W′ 생성
340
W′ 및 V를 기반으로 Q′ 생성
350
종료

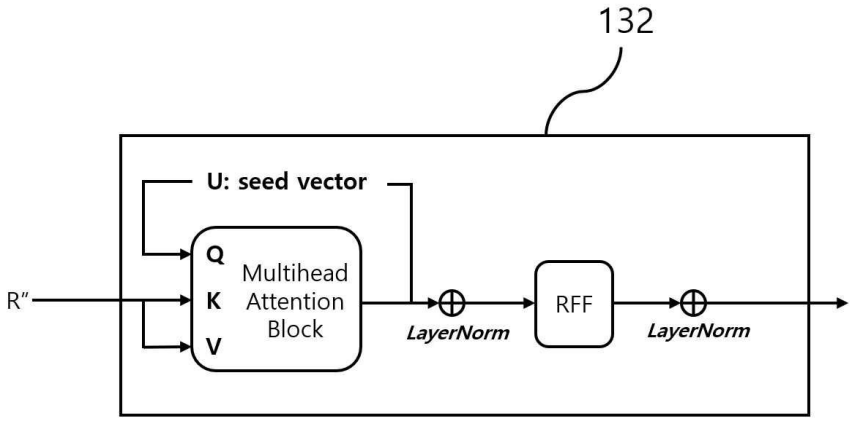
도면4


d
m
아미노산 잔기들의 집합 R
(m x d)
W_q
Query Projection (m x d/4)
모든 query-key 표현형 벡터 쌍에 대해서 어텐션 값 계산
어텐션 값 계산
10 50 40
12 9 9
33 22 55
800 100 100
W (m x n+1)
softmax 정규화
0.1 0.5 0.4
0.4 0.3 0.3
0.3 0.2 0.5
0.8 0.1 0.1
W′ (m x n+1)
n
1
화학적 부분구조들의 집합 S
+ 가상 부분구조 p = S′
(n+1 x d)
W_k
Key Projection (n+1 x d/4)
W_v
Value Projection (n+1 x d/4)
Q′ (m x d/4)

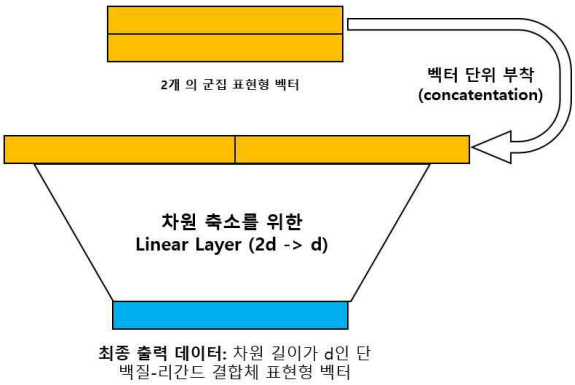
도면5


Head #4
Head #3
Head #2
Head #1
Head #1
Head #2
Head #3
Head #4
O (m x d)

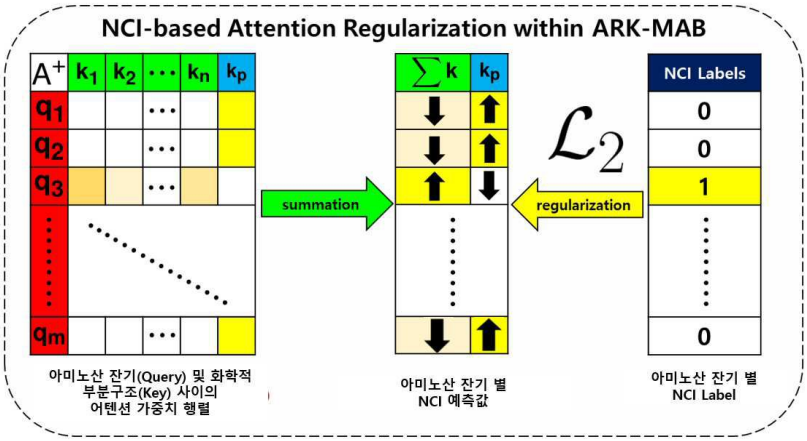
도면6


132
U: seed vector
Q
K
V
Multihead Attention Block
R''
LayerNorm
RFF
LayerNorm

도면7


2개 의 군집 표현형 벡터
벡터 단위 부착 (concatentation)
차원 축소를 위한 Linear Layer (2d -> d)
최종 출력 데이터: 차원 길이가 d인 단백질-리간드 결합체 표현형 벡터

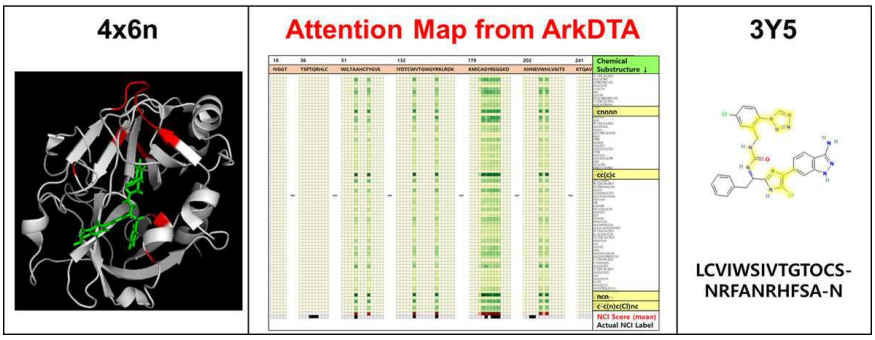
도면8


NCI-based Attention Regularization within ARK-MAB
summation
regularization
NCI Labels
0
0
1
0
아미노산 잔기(Query) 및 화학적 부분구조(Key) 사이의 어텐션 가중치 행렬
아미노산 잔기 별 NCI 예측값
아미노산 잔기 별 NCI Label

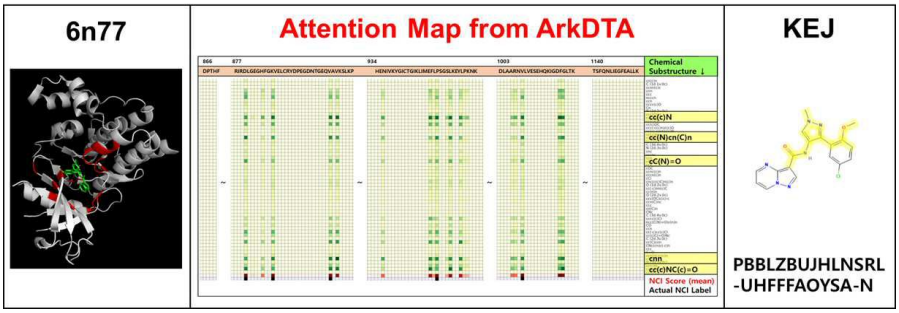
도면9

Model	PDBbind KIKD				PDBbind KIKD (Unseen Scaffolds)				PDBbind IC50				PDBbind IC50 (Unseen Scaffolds)			
	RMSE	MAE	PCORR	CI	RMSE	MAE	PCORR	CI	RMSE	MAE	PCORR	CI	RMSE	MAE	PCORR	CI
ArkDTA (ours)	1.2941 (0.0673)	0.9618 (0.0375)	0.7671 (0.0194)	0.7898 (0.0073)	1.3452 (0.0734)	1.0231 (0.0440)	0.7362 (0.0227)	0.7717 (0.0108)	1.2073 (0.0522)	0.8536 (0.0288)	**0.7752 (0.0169)**	**0.7366 (0.0040)**	**1.1689 (0.0586)**	**0.8368 (0.0340)**	**0.7406 (0.0249)**	**0.7122 (0.0119)**
ArkDTA (Remove $\mathcal{L}_2$)	1.3005 (0.0650)	0.9738 (0.0459)	0.7656 (0.0210)	0.7887 (0.0089)	1.3561 (0.0654)	1.0413 (0.0468)	0.7335 (0.0169)	0.7700 (0.0074)	**1.1931 (0.0360)**	0.8667 (0.0330)	0.7791 (0.0118)	0.7368 (0.0104)	1.1899 (0.0564)	0.8800 (0.0463)	0.7306 (0.0273)	0.7055 (0.0226)
ArkDTA (Freeze ESM)	1.3457 (0.0396)	1.0170 (0.0222)	0.7512 (0.0089)	0.7806 (0.0024)	1.4059 (0.0645)	1.0905 (0.0446)	0.7147 (0.0215)	0.7601 (0.0086)	1.2561 (0.0415)	0.9150 (0.0297)	0.7540 (0.0165)	0.7251 (0.0065)	1.2401 (0.0326)	0.9114 (0.0257)	0.7066 (0.0146)	0.6965 (0.0092)
DeepDTA	1.3589 (0.0534)	0.9761 (0.0405)	0.7657 (0.0145)	0.7879 (0.0077)	1.3651 (0.0338)	1.0354 (0.0517)	0.7278 (0.0225)	0.7687 (0.0119)	1.2131 (0.0468)	0.8794 (0.0291)	0.7687 (0.0259)	0.7349 (0.0124)	1.1963 (0.0655)	0.8721 (0.0439)	0.7195 (0.0453)	0.7071 (0.0160)
GraphDTA	1.5511 (0.0401)	1.1624 (0.0371)	0.6641 (0.0123)	0.7396 (0.0057)	1.5934 (0.0579)	1.2085 (0.0305)	0.6268 (0.0271)	0.7234 (0.0104)	1.5515 (0.1243)	1.1449 (0.1156)	0.6557 (0.0432)	0.6967 (0.0177)	1.5438 (0.1372)	1.1442 (0.1026)	0.5969 (0.0577)	0.6729 (0.0208)
TransformerCPI	1.4982 (0.0375)	1.1478 (0.0256)	0.6716 (0.0171)	0.7455 (0.0079)	1.5008 (0.0551)	1.1606 (0.0356)	0.6519 (0.0236)	0.7348 (0.0090)	1.3465 (0.0332)	0.9646 (0.0373)	0.7190 (0.0120)	0.7144 (0.0083)	1.2905 (0.0761)	0.9332 (0.0538)	0.6816 (0.0399)	0.6931 (0.0164)
MONN	1.3283 (0.0440)	0.9927 (0.0193)	0.7563 (0.0137)	0.7849 (0.0061)	1.3534 (0.0822)	1.0288 (0.0482)	0.7343 (0.0264)	0.7736 (0.0105)	1.3994 (0.0417)	1.0450 (0.0290)	0.6725 (0.0286)	0.6936 (0.0131)	1.3242 (0.0388)	0.9749 (0.0171)	0.6414 (0.0270)	0.6732 (0.0113)
HyperAttentionDTI	1.4028 (0.0186)	1.0888 (0.0192)	0.7257 (0.0149)	0.7683 (0.0061)	1.4480 (0.0418)	1.1399 (0.0351)	0.6899 (0.0153)	0.7510 (0.0067)	1.2994 (0.0293)	0.9816 (0.0265)	0.7371 (0.0142)	0.7233 (0.0047)	1.2652 (0.0406)	0.9594 (0.0332)	0.6879 (0.0295)	0.6962 (0.0049)
AttentionDTA	**1.2711 (0.0454)**	**0.9410 (0.0297)**	**0.7726 (0.0125)**	**0.7908 (0.0063)**	**1.3213 (0.0500)**	**1.0074 (0.0341)**	**0.7372 (0.0169)**	**0.7716 (0.0063)**	1.2131 (0.0639)	**0.8420 (0.0425)**	0.7721 (0.0254)	0.7446 (0.0136)	1.2071 (0.0712)	0.8456 (0.0437)	0.7211 (0.0369)	0.7151 (0.0171)
BACPI	1.3594 (0.0362)	1.0164 (0.0152)	0.7427 (0.0139)	0.7772 (0.0055)	1.4343 (0.0660)	1.0998 (0.0380)	0.6974 (0.0320)	0.7545 (0.0116)	1.3583 (0.0577)	1.0439 (0.0395)	0.7140 (0.0285)	0.7278 (0.0152)	1.3798 (0.0652)	1.0702 (0.0505)	0.6507 (0.0385)	0.7001 (0.0168)
IIFDTI	1.3832 (0.0590)	1.0513 (0.0594)	0.7332 (0.0242)	0.7701 (0.0123)	1.4340 (0.0817)	1.1169 (0.0662)	0.6891 (0.0383)	0.7474 (0.0166)	1.3027 (0.0554)	0.9326 (0.0436)	0.7345 (0.0235)	0.7116 (0.0173)	1.2868 (0.0625)	0.9261 (0.0446)	0.6774 (0.0380)	0.6827 (0.0179)

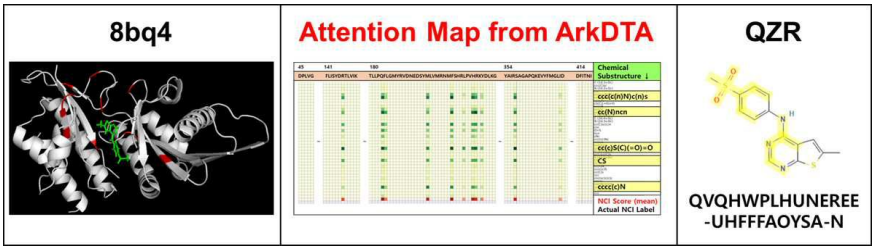
도면10



도면11



도면12



【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구항 1

【변경전】

비공유결합 기반 단백질-리간드 결합친화도 예측 장치에 있어서,

입력된 단백질을 기반으로 m 개의 아미노산 잔기 단위의 표현형 벡터를 생성하는 단백질 인코딩 모듈;

입력된 리간드를 기반으로 n 개의 부분구조 단위의 표현형 벡터를 생성하는 리간드 인코딩 모듈;

상기 m 개의 아미노산 잔기 단위의 표현형 벡터 및 상기 n 개의 부분구조 단위의 표현형 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 백터를 생성하는 단백질-리간드 결합 모듈; 및

상기 단백질-리간드 결합체에 대한 단일 표현형 벡터를 기반으로 상기 단백질과 상기 리간드의 결합친화도를 예측하는 결합친화도 예측 모듈;

을 포함하는 장치에 있어서,

상기 단백질-리간드 결합 모듈은,

상기 m 개의 아미노산 잔기 단위의 표현형 벡터, 상기 n 개의 부분구조 단위의 표현형 벡터 및 학습 가능한 하나의 가상 부분구조 표현형 벡터를 기반으로 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터를 생성하는 ARK-MAB(Attention Regularization based on non-covalent interactions in multihead attention block) 모듈; 및

상기 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터 및 시드 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 벡터를 생성하는 PMA(Pooling by multihead attention layer) 모듈을 포함하며,

상기 ARK-MAB 모듈은,

상기 m 개의 아미노산 잔기 단위의 표현형 벡터를 기반으로 쿼리 프로젝션(query projection) 생성하고,

상기 n 개의 부분구조 단위의 표현형 벡터 및 상기 하나의 가상 부분구조 표현형 벡터를 기반으로 키 프로젝션(key projection) 및 밸류 프로젝션(value projection)을 생성하는 것을 특징으로 하는 장치.

【변경후】

비공유결합 기반 단백질-리간드 결합친화도 예측 장치에 있어서,

입력된 단백질을 기반으로 m 개의 아미노산 잔기 단위의 표현형 벡터를 생성하는 단백질 인코딩 모듈;

입력된 리간드를 기반으로 n 개의 부분구조 단위의 표현형 벡터를 생성하는 리간드 인코딩 모듈;

상기 m 개의 아미노산 잔기 단위의 표현형 벡터 및 상기 n 개의 부분구조 단위의 표현형 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 벡터를 생성하는 단백질-리간드 결합 모듈; 및

상기 단백질-리간드 결합체에 대한 단일 표현형 벡터를 기반으로 상기 단백질과 상기 리간드의 결합친화도를 예측하는 결합친화도 예측 모듈;

을 포함하는 장치에 있어서,

상기 단백질-리간드 결합 모듈은,

상기 m 개의 아미노산 잔기 단위의 표현형 벡터, 상기 n 개의 부분구조 단위의 표현형 벡터 및 학습 가능한 하나의 가상 부분구조 표현형 벡터를 기반으로 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터를 생성하는 ARK-MAB(Attention Regularization based on non-covalent interactions in multihead attention block) 모듈; 및

상기 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터 및 시드 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 벡터를 생성하는 PMA(Pooling by multihead attention layer) 모듈을 포함하며,

상기 ARK-MAB 모듈은,

상기 m 개의 아미노산 잔기 단위의 표현형 벡터를 기반으로 쿼리 프로젝션(query projection)을 생성하고,

상기 n 개의 부분구조 단위의 표현형 벡터 및 상기 하나의 가상 부분구조 표현형 벡터를 기반으로 키 프로젝션(key projection) 및 밸류 프로젝션(value projection)을 생성하는 것을 특징으로 하는 장치.

【직권보정 2】

【보정항목】 청구범위

【보정세부항목】 청구항 9

【변경전】

비공유결합 기반 단백질-리간드 결합친화도 예측 방법에 있어서,

(a) 단백질 인코딩 모듈이 입력된 단백질을 기반으로 m 개의 아미노산 잔기 단위의 표현형 벡터를 생성하는 단계;

(b) 리간드 인코딩 모듈이 입력된 리간드를 기반으로 n 개의 부분구조 단위의 표현형 벡터를 생성하는 단계;

(c) 단백질-리간드 결합 모듈이 상기 m 개의 아미노산 잔기 단위의 표현형 벡터 및 상기 n 개의 부분구조 단위의 표현형 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 백터를 생성하는 단계; 및

(d) 결합친화도 예측 모듈이 상기 단백질-리간드 결합체에 대한 단일 표현형 벡터를 기반으로 상기 단백질과 상기 리간드의 결합친화도를 예측하는 단계;

를 포함하는 방법에 있어서,

상기 (c) 단계는,

(c-1) 상기 m 개의 아미노산 잔기 단위의 표현형 벡터, 상기 n 개의 부분구조 단위의 표현형 벡터 및 학습 가능한 하나의 가상 부분구조 표현형 벡터를 기반으로 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터를 생성하는 단계; 및

(c-2) 상기 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터 및 시드 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 벡터를 생성하는 단계;

를 포함하며,

(c-3) 상기 m 개의 아미노산 잔기 단위의 표현형 벡터를 기반으로 쿼리 프로젝션(query projection)을 생성하는 단계; 및,

(c-4) 상기 n 개의 부분구조 단위의 표현형 벡터 및 상기 하나의 가상 부분구조 표현형 벡터를 기반으로 키 프로젝션(key projection) 및 밸류 프로젝션(value projection)을 생성하는 단계;

를 더 포함하는 것을 특징으로 하는 방법.

【변경후】

비공유결합 기반 단백질-리간드 결합친화도 예측 방법에 있어서,

(a) 단백질 인코딩 모듈이 입력된 단백질을 기반으로 m 개의 아미노산 잔기 단위의 표현형 벡터를 생성하는 단계;

(b) 리간드 인코딩 모듈이 입력된 리간드를 기반으로 n 개의 부분구조 단위의 표현형 벡터를 생성하는 단계;

(c) 단백질-리간드 결합 모듈이 상기 m 개의 아미노산 잔기 단위의 표현형 벡터 및 상기 n 개의 부분구조 단위의 표현형 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 벡터를 생성하는 단계; 및

(d) 결합친화도 예측 모듈이 상기 단백질-리간드 결합체에 대한 단일 표현형 벡터를 기반으로 상기 단백질과 상기 리간드의 결합친화도를 예측하는 단계;

를 포함하는 방법에 있어서,

상기 (c) 단계는,

(c-1) 상기 m 개의 아미노산 잔기 단위의 표현형 벡터, 상기 n 개의 부분구조 단위의 표현형 벡터 및 학습 가능한 하나의 가상 부분구조 표현형 벡터를 기반으로 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터를 생성하는 단계; 및

(c-2) 상기 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터 및 시드 벡터를 기반으로 단백질-리간드 결합체에

대한 단일 표현형 벡터를 생성하는 단계;

를 포함하며,

(c-3) 상기 m 개의 아미노산 잔기 단위의 표현형 벡터를 기반으로 쿼리 프로젝션(query projection)을 생성하는 단계; 및,

(c-4) 상기 n 개의 부분구조 단위의 표현형 벡터 및 상기 하나의 가상 부분구조 표현형 벡터를 기반으로 키 프로젝션(key projection) 및 밸류 프로젝션(value projection)을 생성하는 단계;

를 더 포함하는 것을 특징으로 하는 방법.

【직권보정 3】

【보정항목】 청구범위

【보정세부항목】 청구항 10

【변경전】

컴퓨팅 장치와 결합하여,

(A) 입력된 단백질을 기반으로 m 개의 아미노산 잔기 단위의 표현형 벡터를 생성하는 단계;

(B) 입력된 리간드를 기반으로 n 개의 부분구조 단위의 표현형 벡터를 생성하는 단계;

(C) 상기 m 개의 아미노산 잔기 단위의 표현형 벡터 및 상기 n 개의 부분구조 단위의 표현형 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 백터를 생성하는 단계; 및

(D) 상기 단백질-리간드 결합체에 대한 단일 표현형 벡터를 기반으로 상기 단백질과 상기 리간드의 결합친화도를 예측하는 단계;

를 실행시키기 위하여 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램에 있어서,

상기 (C) 단계는,

(C-1) 상기 m 개의 아미노산 잔기 단위의 표현형 벡터, 상기 n 개의 부분구조 단위의 표현형 벡터 및 학습 가능한 하나의 가상 부분구조 표현형 벡터를 기반으로 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터를 생성하는 단계; 및

(C-2) 상기 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터 및 시드 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 벡터를 생성하는 단계;

를 포함하며,

(C-3) 상기 m 개의 아미노산 잔기 단위의 표현형 벡터를 기반으로 쿼리 프로젝션(query projection)을 생성하는 단계; 및,

(C-4) 상기 n 개의 부분구조 단위의 표현형 벡터 및 상기 하나의 가상 부분구조 표현형 벡터를 기반으로 키 프로젝션(key projection) 및 밸류 프로젝션(value projection)을 생성하는 단계;

를 더 포함하는 것을 특징으로 하는 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램.

【변경후】

컴퓨팅 장치와 결합하여,

(A) 입력된 단백질을 기반으로 m 개의 아미노산 잔기 단위의 표현형 벡터를 생성하는 단계;

(B) 입력된 리간드를 기반으로 n 개의 부분구조 단위의 표현형 벡터를 생성하는 단계;

(C) 상기 m 개의 아미노산 잔기 단위의 표현형 벡터 및 상기 n 개의 부분구조 단위의 표현형 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 벡터를 생성하는 단계; 및

(D) 상기 단백질-리간드 결합체에 대한 단일 표현형 벡터를 기반으로 상기 단백질과 상기 리간드의 결합친화도를 예측하는 단계;

를 실행시키기 위하여 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램에 있어서,

상기 (C) 단계는,

(C-1) 상기 m 개의 아미노산 잔기 단위의 표현형 벡터, 상기 n 개의 부분구조 단위의 표현형 벡터 및 학습 가능한 하나의 가상 부분구조 표현형 벡터를 기반으로 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터를 생성하는 단계; 및

(C-2) 상기 m 개의 정제된 아미노산 잔기 단위의 표현형 벡터 및 시드 벡터를 기반으로 단백질-리간드 결합체에 대한 단일 표현형 벡터를 생성하는 단계;

를 포함하며,

(C-3) 상기 m 개의 아미노산 잔기 단위의 표현형 벡터를 기반으로 쿼리 프로젝션(query projection)을 생성하는 단계; 및,

(C-4) 상기 n 개의 부분구조 단위의 표현형 벡터 및 상기 하나의 가상 부분구조 표현형 벡터를 기반으로 키 프로젝션(key projection) 및 밸류 프로젝션(value projection)을 생성하는 단계;

를 더 포함하는 것을 특징으로 하는 컴퓨터로 판독 가능한 매체에 저장된 컴퓨터 프로그램.