(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년12월27일
(11) 등록번호 10-2617958
(24) 등록일자 2023년12월20일

(51) 국제특허분류(Int. Cl.)
G16C 20/70 (2019.01) *G06N 3/0464* (2023.01)
G16C 20/20 (2019.01) *G16C 20/30* (2019.01)
G16C 20/80 (2019.01) *G16C 20/90* (2019.01)
(52) CPC특허분류
G16C 20/70 (2019.02)
G06N 3/0464 (2023.01)
(21) 출원번호 10-2022-0184080
(22) 출원일자 2022년12월26일
심사청구일자 2022년12월26일
(56) 선행기술조사문헌
B. Chen, Molecular Graph Representation Learning and Generation for Drug Discovery, Thesis, Massachusetts Institute of Technology(2022.06.15.)*
H. Kurata et al., ICAN: Interpretable cross-attention network for identifying drug and target protein interactions, PLoS ONE 17, e0276609(2022.10.24.)*
L. Simone, Learning Text Embeddings on Virus Sequences for Candidate Drug Discovery, Thesis, Politecnico di Milano(2020.12.15.)*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
고려대학교산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
강재우
서울특별시 성동구 왕십리로 241 서울숲더샵, 103동 1402호
장광훈
서울특별시 동대문구 제기로2가길 3, 203호
응웬옥광
인천광역시 연수구 비류대로519번길 201동 103호
(74) 대리인
특허법인위솔

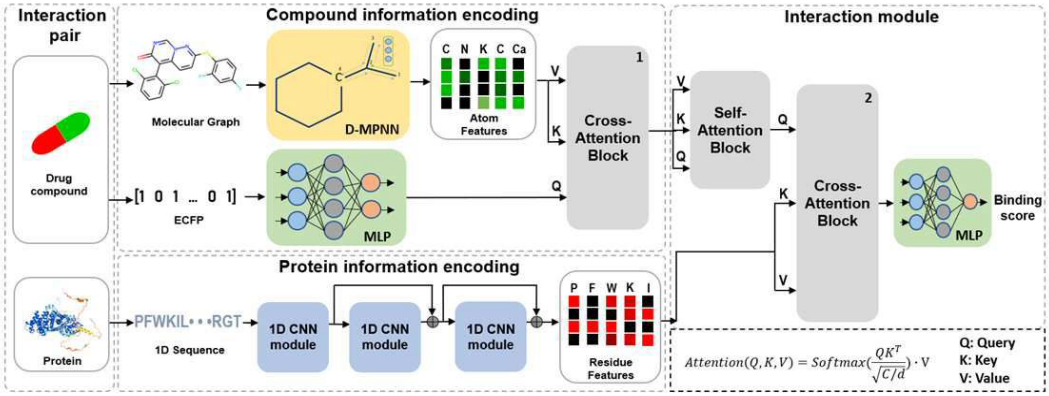
전체 청구항 수 : 총 8 항 심사관 : 권계민

(54) 발명의 명칭 **교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법 및 장치**

(57) 요 약

본 발명은 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법 및 장치에 관한 것이다. 본 발명의 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법은, 분자 그래프 데이터 및 분자 지문 데이터를 기반으로 화합물 정보를 인코딩하는 단계, 단백질 시퀀스 데이터를 기반으로 단백질 정보를 인코딩하는 단계, 인코딩된 상기 화합물 정보 및 단백질 정보를 제1 교차 어텐션 블록에 입력하는 단계 및 상기 제1 교차 어텐션 블록의 출력을 기반으로 화합물과 단백질의 상호 작용을 예측하는 단계를 포함할 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G16C 20/20 (2019.02)
G16C 20/30 (2019.02)
G16C 20/80 (2019.02)
G16C 20/90 (2019.02)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711168358
과제번호 2020R1A2C3010638
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 개인기초연구(과기정통부)
연구과제명 생명의료 빅데이터와 딥러닝 기반 데이터주도 신약후보물질 발굴
기 여 율 1/3
과제수행기관명 고려대학교
연구기간 2022.03.01 ~ 2023.02.28
이 발명을 지원한 국가연구개발사업
과제고유번호 1465036037
과제번호 HR20C0021030022
부처명 보건복지부
과제관리(전문)기관명 한국보건산업진흥원
연구사업명 연구중심병원 육성 R&D
연구과제명 인공지능·빅데이터 기반 데이터주도 의료기술 개발 및 고도화
기 여 율 1/3
과제수행기관명 고려대학교산학협력단
연구기간 2022.01.01 ~ 2022.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711159611
과제번호 2020-0-01819-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 ICT명품인재양성(고려대학교)
기 여 율 1/3
과제수행기관명 고려대학교산학협력단
연구기간 2022.01.01 ~ 2022.12.31
공지예외적용 : 있음

명 세 서

청구범위

청구항 1

교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법에 있어서,

분자 그래프 데이터 및 분자 지문 데이터 각각을 기반으로 화합물 정보를 인코딩하는 단계;

단백질 시퀀스 데이터를 기반으로 단백질 정보를 인코딩하는 단계;

인코딩된 상기 화합물 정보 및 단백질 정보를 제1 교차 어텐션 블록에 입력하는 단계; 및

상기 제1 교차 어텐션 블록의 출력을 기반으로 화합물과 단백질의 상호 작용을 예측하는 단계를 포함하는 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법에 있어서,

상기 제1 교차 어텐션 블록은 Q(query), K(Key) 및 V(value)의 입력을 가지고, 상기 Q는 인코딩된 상기 화합물 정보를 기반으로 결정되고, 상기 K 및 V는 인코딩된 상기 단백질 정보를 기반으로 결정되며,

상기 화합물 정보를 인코딩하는 단계는,

상기 분자 그래프 데이터를 D-MPNN(direct message passing neural network)에 입력하는 단계;

상기 분자 지문 데이터를 MLP(multi layer perceptron) 인공 신경망에 입력하는 단계; 및

상기 D-MPNN의 출력 및 상기 MLP 인공 신경망의 출력을 제2 교차 어텐션 블록에 입력하여 인코딩된 상기 화합물 정보이자 상기 제1 교차 어텐션 블록에 입력되는 Q를 출력하는 단계;

를 포함하는,

교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법.

청구항 2

삭제

청구항 3

제1항에 있어서,

상기 제2 교차 어텐션 블록은 Q(query), K(Key) 및 V(value)의 입력을 가지고, 상기 V, K는 상기 D-MPNN의 출력을 기반으로 결정되고, 상기 Q는 상기 MLP 인공 신경망의 출력을 기반으로 결정되는 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법.

청구항 4

제3항에 있어서,

상기 제2 교차 어텐션 블록의 출력은, 상기 Q, K, V의 입력을 기반으로

$$CrossAttention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{C/d}}\right) * V$$

와 같이 결정되고, 여기서 C는 임베딩 차원의 수이고, d는 어텐션 헤드(attention head)의 수인 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법.

청구항 5

제1항에 있어서,

상기 화합물 정보를 인코딩하는 단계는,

상기 제2 교차 어텐션 블록의 출력을 셀프 어텐션 블록에 입력하는 단계를 더 포함하는 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법.

청구항 6

제1항에 있어서,

단백질 정보를 인코딩하는 단계는,

TAPE(tasks assessing protein embeddings) 토크나이저(tokenizer)를 기반으로 단백질 데이터를 전처리하는 단계;

One-hot 인코딩을 기반으로 1 차원 데이터인 상기 단백질 시퀀스 데이터를 생성하는 단계; 및

상기 단백질 시퀀스 데이터를 1D CNN(convolutional neural network)에 입력하는 단계를 포함하는 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법.

청구항 7

삭제

청구항 8

제1항에 있어서,

상기 제1 교차 어텐션 블록의 출력은, 상기 Q, K, V의 입력을 기반으로

$$CrossAttention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{C/d}}\right) * V$$

와 같이 결정되고, 여기서 C는 임베딩 차원의 수이고, d는 어텐션 헤드(attention head)의 수인 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법.

청구항 9

적어도 하나의 프로세서;

상기 적어도 하나의 프로세서에 의해 실행되는 컴퓨터 프로그램을 로드(load)하는 메모리; 및

상기 컴퓨터 프로그램을 저장하는 스토리지를 포함하고,

상기 컴퓨터 프로그램은 상기 적어도 하나의 프로세서에 의해,

분자 그래프 데이터 및 분자 지문 데이터 각각을 기반으로 화합물 정보를 인코딩하는 동작;

단백질 시퀀스 데이터를 기반으로 단백질 정보를 인코딩하는 동작;

인코딩된 상기 화합물 정보 및 단백질 정보를 제1 교차 어텐션 블록에 입력하는 동작; 및

상기 제1 교차 어텐션 블록의 출력을 기반으로 화합물과 단백질의 상호 작용을 예측하는 동작을 실행하도록 제어되는 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 장치에 있어서,

상기 제1 교차 어텐션 블록은 Q(query), K(Key) 및 V(value)의 입력을 가지고, 상기 Q는 인코딩된 상기 화합물 정보를 기반으로 결정되고, 상기 K 및 V는 인코딩된 상기 단백질 정보를 기반으로 결정되며,

상기 화합물 정보를 인코딩하는 동작은,

상기 분자 그래프 데이터를 D-MPNN(direct message passing neural network)에 입력하는 동작;

상기 분자 지문 데이터를 MLP(multi layer perceptron) 인공 신경망에 입력하는 동작; 및

상기 D-MPNN의 출력 및 상기 MLP 인공 신경망의 출력을 제2 교차 어텐션 블록에 입력하여 인코딩된 상기 화합물 정보이자 상기 제1 교차 어텐션 블록에 입력되는 Q를 출력하는 동작;

을 포함하는,

교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 장치.

청구항 10

컴퓨팅 장치와 결합되어,

분자 그래프 데이터 및 분자 지문 데이터 각각을 기반으로 화합물 정보를 인코딩하는 동작;

단백질 시퀀스 데이터를 기반으로 단백질 정보를 인코딩하는 동작;

인코딩된 상기 화합물 정보 및 단백질 정보를 제1 교차 어텐션 블록에 입력하는 동작; 및

상기 제1 교차 어텐션 블록의 출력을 기반으로 화합물과 단백질의 상호 작용을 예측하는 동작을 실행하는 컴퓨터로 판독 가능한 기록 매체에 저장된 컴퓨터 프로그램에 있어서,

상기 제1 교차 어텐션 블록은 Q(query), K(Key) 및 V(value)의 입력을 가지고, 상기 Q는 인코딩된 상기 화합물 정보를 기반으로 결정되고, 상기 K 및 V는 인코딩된 상기 단백질 정보를 기반으로 결정되며,

상기 화합물 정보를 인코딩하는 동작은,

상기 분자 그래프 데이터를 D-MPNN(direct message passing neural network)에 입력하는 동작;

상기 분자 지문 데이터를 MLP(multi layer perceptron) 인공 신경망에 입력하는 동작; 및

상기 D-MPNN의 출력 및 상기 MLP 인공 신경망의 출력을 제2 교차 어텐션 블록에 입력하여 인코딩된 상기 화합물 정보이자 상기 제1 교차 어텐션 블록에 입력되는 Q를 출력하는 동작;

을 포함하는,

컴퓨터로 판독 가능한 기록 매체에 저장된 컴퓨터 프로그램.

발명의 설명

기 술 분 야

[0001] 본 발명은 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법 및 장치에 관한 것이다. 보다 상세하게는, 그래프 컨볼로션 신경망과 분자 지문(Molecular fingerprints) 데이터를 동시에 사용하면서 화합물과 단백질 정보의 단순 결합이 아닌 교차 어텐션 매커니즘을 이용한 화합물-단백질 상호작용 예측 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 화합물-단백질 상호작용 예측은 신약 개발과정에서 필수적인 단계로 비용이 높은 분자 도킹 시뮬레이션을 통해 수행된다. 이를 대체하기 위해 인공 지능 기반의 접근 방식들이 제안되고 있다. 최근 분자 구조를 그래프 데이터로 사용해 학습하는 그래프 컨볼루션 신경망 모델과 분자의 부분 구조를 비트 단위로 치환한 분자 지문(Molecular fingerprints) 데이터를 학습하는 인공 신경망 모델로 크게 두 가지 딥러닝 모델이 화합물 데이터를 인공 지능에 활용하는데 사용되고 있다. 대부분의 기존 연구는 두 방법론 중 하나만을 사용한다. 또한 최근 연구는 화합물-단백질 상호작용 예측 연구를 위해 화합물 및 단백질 데이터를 단순히 결합하여 학습하는 방법을 넘어 학습 시 데이터들간의 결합을 어떻게 모델링할지에 대한 연구가 활발히 진행되고 있다. 본 발명은 그래프 컨볼로션 신경망과 분자 지문 데이터를 동시에 사용하면서 화합물과 단백질 정보의 단순 결합이 아닌 교차 어텐션 매커니즘을 이용한 화합물-단백질 상호작용 예측 방법 및 장치에 관한 것이다.

선행기술문헌

특허문헌

[0003] (특허문헌 0001) 대한민국 공개특허공보 제10-2020-0124923호(2020.11.04)

발명의 내용

해결하려는 과제

[0004] 본 발명은 분자 구조 그래프 데이터 혹은 지문 데이터 중 어느 하나만을 사용하는 기존 기술과 비교해서 효과적이고 개선된 성능을 가지는 화합물-단백질 상호작용 예측 방법을 제안하는 것을 목적으로 한다.

[0005] 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

[0006] 전술한 기술적 과제를 달성하기 위한 본 발명의 일 실시 예에 따른 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법은, 분자 그래프 데이터 및 분자 지문 데이터를 기반으로 화합물 정보를 인코딩하는 단계, 단백질 시퀀스 데이터를 기반으로 단백질 정보를 인코딩하는 단계, 인코딩된 상기 화합물 정보 및 단백질 정보를 제1 교차 어텐션 블록에 입력하는 단계 및 상기 제1 교차 어텐션 블록의 출력을 기반으로 화합물과 단백질의 상호 작용을 예측하는 단계를 포함할 수 있다.

[0007] 일 실시 예에 따르면, 상기 화합물 정보를 인코딩하는 단계는, 상기 분자 그래프 데이터를 D-MPNN(direct message passing neural network)에 입력하는 단계, 상기 분자 지문 데이터를 MLP(mulit layer perceptron) 인공 신경망에 입력하는 단계 및 상기 D-MPNN의 출력 및 상기 MLP 인공 신경망의 출력을 제2 교차 어텐션 블록에 입력하는 단계를 포함할 수 있다.

[0008] 일 실시 예에 따르면, 상기 제2 교차 어텐션 블록은 Q(query), K(Key), V(value)의 입력을 가지고, 상기 V, K는 상기 D-MPNN의 출력을 기반으로 결정되고, 상기 Q는 상기 MLP 인공 신경망의 출력을 기반으로 결정될 수 있다.

[0009] 일 실시 예에 따르면, 상기 제2 교차 어텐션 블록의 출력은, 상기 Q, K, V의 입력을 기반으로, $CrossAttention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{C/d}}\right) * V$ 와 같이 결정되고, 여기서 C는 임베딩 차원의 수이고, d는 어텐션 헤드(attention head)의 수일 수 있다.

[0010] 일 실시 예에 따르면, 상기 화합물 정보를 인코딩하는 단계는, 상기 제2 교차 어텐션 블록의 출력을 셀프 어텐션 블록에 입력하는 단계를 더 포함할 수 있다.

[0011] 일 실시 예에 따르면, 단백질 정보를 인코딩하는 단계는, TAPE(tasks assessing protein embeddings) 토크나이저(tokenizer)를 기반으로 단백질 데이터를 전처리하는 단계, One-hot 인코딩을 기반으로 1 차원 데이터인 상기 단백질 시퀀스 데이터를 생성하는 단계 및 상기 단백질 시퀀스 데이터를 1D CNN(convolutional neural network)에 입력하는 단계를 포함할 수 있다.

[0012] 일 실시 예에 따르면, 상기 제1 교차 어텐션 블록은 Q(query), K(Key), V(value)의 입력을 가지고, 상기 Q는 인코딩된 상기 화합물 정보를 기반으로 결정되고, 상기 K 및 V는 인코딩된 상기 단백질 정보를 기반으로 결정될 수 있다.

[0013] 일 실시 예에 따르면, 상기 제1 교차 어텐션 블록의 출력은, 상기 Q, K, V의 입력을 기반으로 $CrossAttention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{C/d}}\right) * V$ 와 같이 결정되고, 여기서 C는 임베딩 차원의 수이고, d는 어텐션 헤드(attention head)의 수일 수 있다.

[0014] 본 발명의 기술적 과제를 달성하기 위한 본 발명의 일 실시 예에 따른 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 장치는, 적어도 하나의 프로세서, 상기 적어도 하나의 프로세서에 의해 실행되는 컴퓨터 프로그램을 로드(load)하는 메모리 및 상기 컴퓨터 프로그램을 저장하는 스토리지를 포함하고, 상기 컴퓨터 프로그램은 상기 적어도 하나의 프로세서에 의해, 분자 그래프 데이터 및 분자 지문 데이터를 기반으로 화합물 정보를 인코딩하는 동작, 단백질 시퀀스 데이터를 기반으로 단백질 정보를 인코딩하는 동작, 인코딩된 상기 화합물 정보 및 단백질 정보를 제1 교차 어텐션 블록에 입력하는 동작 및 상기 제1 교차 어텐션 블록의 출력을 기반으로 화합물과 단백질의 상호 작용을 예측하는 동작을 실행하도록 제어될 수 있다.

[0015] 본 발명의 기술적 과제를 달성하기 위한 본 발명의 일 실시 예에 따른 컴퓨터 판독 가능한 기록 매체에 저장된 컴퓨터 프로그램은 컴퓨팅 장치와 결합되어, 분자 그래프 데이터 및 분자 지문 데이터를 기반으로 화합물 정보

를 인코딩하는 동작, 단백질 시퀀스 데이터를 기반으로 단백질 정보를 인코딩하는 동작, 인코딩된 상기 화합물 정보 및 단백질 정보를 제1 교차 어텐션 블록에 입력하는 동작 및 상기 제1 교차 어텐션 블록의 출력을 기반으로 화합물과 단백질의 상호 작용을 예측하는 동작을 실행할 수 있다.

발명의 효과

[0016] 본 발명에 따르면, 분자 구조 그래프 데이터 혹은 지문 데이터 중 어느 하나만을 사용하는 기존 기술과 비교해서 효과적이고 개선된 성능을 가지는 화합물-단백질 상호작용 예측 방법이 개시된다.

[0017] 본 발명의 효과들은 이상에서 언급한 효과들로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해 될 수 있을 것이다.

도면의 간단한 설명

[0018] 도 1은 본 발명의 일 실시 예에 따른 화합물-단백질 상호작용 예측 모델의 구조를 나타낸 도면이다.

도 2는 본 발명의 일 실시 예에 따른 교차 어텐션 매카니즘의 세부 구조를 나타낸 도면이다.

도 3은 본 발명의 일 실시 예에 따른 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법을 나타낸 순서도이다.

도 4는 본 발명의 일 실시 예에 따른 화합물 정보를 인코딩하는 과정을 나타낸 도면이다.

도 5는 본 발명의 일 실시 예에 따른 단백질 정보를 인코딩하는 과정을 나타낸 도면이다.

도 6는 본 발명의 일 실시 예에 따른 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 장치의 구성을 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0019] 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.

[0020] 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.

[0021] 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.

[0022] 또한, 어떤 구성요소들을 포함한다는 표현은 "개방형"의 표현으로서 해당 구성요소들이 존재하는 것을 단순히 지칭할 뿐이며, 추가적인 구성요소들을 배제하는 것으로 이해되어서는 안 된다.

[0023] 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.

[0024] 이하에서는 도면들을 참조하여 본 발명의 세부적인 실시 예들에 대해 살펴보도록 한다.

[0025] 도 1은 본 발명의 일 실시 예에 따른 화합물-단백질 상호작용 예측 모델의 구조를 나타낸 도면이다.

[0026] 도 1을 참조하면, 본 발명의 일 실시 예에 따른 화합물-단백질 상호작용 예측 모델은 D-MPNN(direct message passing neural network), multi layer perceptrons (MLP) 인공 신경망, 1D CNN (convolutional neural network)을 포함하는 3 개의 입력 인공 신경망, 어텐션 블록(attention block), 그리고 출력 인공 신경망으로 구성될 수 있다. D-MPNN, MLP 인공 신경망은 화합물 정보를 인코딩하기 위해 사용되며, 1D CNN은 단백질 정보를 인코딩하기 위해 사용될 수 있다. D-MPNN, MLP, 그리고 1D CNN을 통해 인코딩된 화합물 정보 및 단백질 정보는

교차 어텐션 블록을 통해 결합되며, 출력 인공 신경망을 통해 예측된 Binding Score가 출력될 수 있다.

[0027] 우선, 입력 신경망을 통해 화합물 정보를 인코딩하는 방법에 대해 설명한다. 도 1을 참조하면, 입력 인공 신경망 중 D-MPNN 및 MLP 인공 신경망은 각각 분자 구조 그래프 데이터와 분자 지문 데이터(Morgan fingerprint 데이터)를 입력받을 수 있다.

[0028] 여기서 분자 구조 그래프 데이터 및 분자 지문 데이터는 아래와 같이 표현될 수 있다.

[0029] - $G_s(A,B)$는 원자들의 집합^{A}와 결합들의 집합^{B}로 구성된 분자 그래프 데이터를 나타낸다.

[0030] - Morgan fingerprint 벡터 M_s는 이원 벡터로 특정 부분 구조의 존재 유무를 0과 1로 표현한다. Morgan 알고리즘은 분자의 모든 원자에 이르는 경로를 전부 탐색하며, Morgan fingerprint 벡터 M_s는 고유 경로를 비트로 치환한 데이터를 나타낸다.

[0031] 분자 그래프 데이터 $G_s(A,B)$는 D-MPNN에 입력되고, Moragn fingerprint 벡터 M_s는 MLP 인공 신경망에 입력될 수 있다. D-MPNN 모델은 분자 구조 그래프의 노드가 아닌 방향성이 있는 간선을 이용해 원자간의 관계를 학습시키는 모델로 O_{G_s}를 출력할 수 있다. MLP 인공 신경망은 입력 M_s로부터 비선형 정보를 추출하여 O_{M_s}를 출력할 수 있다.

[0032] 두 출력 값 O_{M_s} 및 O_{G_s}는 교차 어텐션 매커니즘을 통해 결합될 수 있다. 교차-어텐션 매커니즘은 Q(Query), K(Key), V(Value)를 입력값으로 가지며, Q 는 MLP 인공 신경망의 출력인 O_{M_s}로부터, 그리고 K,V는 D-MPNN의 출력인 O_{G_s}로부터 생성될 수 있다. 이를 수학식으로 나타내면 수학식 1과 같다.

수학식 1

[0033] $$Q = f_Q(O_{M_s});\ K = f_k(O_{G_s});\ V = f_v(O_{G_s})$$

[0034] 수학식 1과 같이, Q 는 O_{M_s}로부터 $Q = f_Q(O_{M_s})$로서 생성되고, K,V는 O_{G_s}로부터 각각 $K = f_k(O_{G_s})$, $V = f_v(O_{G_s})$로서 생성될 수 있다. 이 때 f는 projection function $f = w^T x + b$ (w와 b는 각각 가중치와 편향을 의미함)와 같이 정의될 수 있다.

[0035] 도 2는 본 발명의 일 실시 예에 따른 교차 어텐션 매카니즘의 세부 구조를 나타낸 도면이다.

[0036] D-MPNN 및 MLP 인공 신경망을 통해 출력된 O_{M_s} 및 O_{G_s}는 Q,K,V를 입력 값으로 가지는 교차-어텐션 매커니즘을 통해 결합될 수 있다. 이를 수학식으로 나타내면 수학식 2와 같다.

수학식 2

[0037] $$Attention_energy = Softmax\left(\frac{QK^T}{\sqrt{C/d}}\right)$$

[0038] $$Comp_s = CrossAttention(Q,K,V) = Attention_{energy} * V$$

[0039] 여기서 C와 d는 각각 임베딩 차원의 수와 attention head의 수를 의미한다. 본 모델은 single-head cross-attention mechanism($d = 1$)을 사용하였으나, 이는 예시를 위한 것일 뿐 본 발명의 범위를 한정하지 않음에 유의한다. 본 발명의 일 실시 예에 따르면, 화합물 표현형을 개선하기 위해 셀프-어텐션(self-attention)을 추가

로 적용시켜 수학식 2와 같이 화합물 Feautre의 최종 표현형 $Comp_s$를 생성할 수 있다.

[0040] 다음으로, 단백질 정보를 인코딩하는 방법에 대해 설명한다. 도 1을 참조하면, 입력 인공 신경망 중 1-D CNN은 단백질 시퀀스 데이터를 입력받을 수 있다. 단백질 데이터는 residue를 UniRep Vocabulary에서 부여한 고유 숫자로 표현할 수 있는 TAPE(Tasks Assessing Protein Embeddings) tokenizer를 이용해 전처리 한 후 One-hot encoding을 이용해 단백질 시퀀스를 1차원 데이터로 구성한 뒤 1-D CNN 모델에 입력될 수 있다. 이를 통해 단백질 Feature인 $Prot_t$가 생성될 수 있다.

[0041] 정확한 화합물-단백질 상호작용을 예측하기 위해선 입력 인공 신경망들을 통해 얻은 화합물 Feature와 단백질 feature들($Comp_s, Prot_t$)을 결합시키는 방법이 중요하다. 본 발명은 화합물의 구조의 다양한 데이터들을 결합한 뒤 최종적으로 단백질 데이터와 교차 어텐션 모듈을 이용해 결합시켜 상호 작용을 예측하는 새로운 비대칭적 인공지능 모델을 제안한다. 교차 어텐션 모듈은 다양한 단백질 데이터에 대해 화합물 데이터가 주는 영향의 패턴을 파악하도록 학습된다. 다시 말해 단백질이 화합물과 반응하는 방식을 모델링한다. 본 모델에서 $Comp_s$와 $Prot_t$의 결합 공식은 아래 수학식 3과 같다.

수학식 3

[0042] $$Q = f_Q(Comp_s);\ K = f_k(Prot_t);\ V = f_v(Prot_t)$$

[0043] $$Interaction_{s,t} = CrossAttention(Q, K, V)$$

[0044] $Interaction_{s,t}$는 출력 인공 신경망에 입력되어 단백질과 화합물간의 Binding score를 생성하며 이를 기반으로 화합물과 단백질 간 상호 작용을 예측할 수 있다. 예를 들어, 교차 어텐션 블록의 출력인 $Interaction_{s,t}$는 MLP 인공 신경망에 입력되며 MLP 인공 신경망은 출력으로 Binding Score를 생성할 수 있다. 해당 Binding Score를 기반으로 화합물과 단백질 간 최종적인 상호 작용을 예측할 수 있다.

[0045] 도 3은 본 발명의 일 실시 예에 따른 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법을 나타낸 순서도이다. 이하 도면을 참조하여 설명할 본 발명의 일 실시 예에 따른 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법의 각 단계는 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 장치에 의해 수행되는 것으로 이해될 수 있다. 또한 이하 본 발명을 설명함에 있어서 별도로 정의하지 않는 한 "장치"는 본 발명의 다양한 실시 예에 따른 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 장치를 의미하는 것으로 이해될 수 있다.

[0046] 도 3을 참조하면, 장치는 우선 분자 그래프 및 분자 지문 데이터(Morgan fingerprint 데이터)를 기반으로 화합물 정보를 인코딩할 수 있다(S310). 분자 그래프 데이터는 원자들의 집합 ^{A}와 결합들의 집합 ^{B}로 구성된 분자 그래프 데이터를 나타내는 $G_s(A,B)$로 표현될 수 있다. 분자 지문 데이터는 이원 벡터로 특정 부분 구조의 존재 유무를 0 과 1로 표현하는 Morgan fingerprint 벡터 M_s로 표현될 수 있다. S310 단계에서 화합물 정보를 인코딩하기 위한 과정을 이하 도 4를 참조하여 설명한다.

[0047] 도 4는 본 발명의 일 실시 예에 따른 화합물 정보를 인코딩하는 과정을 나타낸 도면이다.

[0048] 도 4를 참조하면, 분자 그래프 데이터 $G_s(A,B)$는 D-MPNN(direct message passing neural network)에 입력될 수 있다(S410). 여기서 D-MPNN은 분자 구조 그래프의 노드가 아닌 방향성이 있는 간선을 이용해 원자 간의 관계를 학습시키는 모델로서 O_{G_s}를 출력할 수 있다. 또한, Morgan fingerprint M_s로 표현되는 벡터 분자 지문 데이터는 MLP(multi layer perceptron) 인공 신경망에 입력될 수 있다(S420). MLP 인공 신경망은 입력 M_s로부터 비선형 정보를 추출하여 O_{M_s}를 출력할 수 있다.

[0049] D-MPNN의 출력 O_{M_s} 및 MLP 인공 신경망의 출력 O_{G_s}는 교차 어텐션 블록에 입력되며, 교차 어텐션 매커니즘을 통해 결합될 수 있다(S430). S430 단계에서 사용되는 교차-어텐션 블록은 Q(Query), K(Key), V(Value)를 입력값으로 가지며, Q 는 MLP 인공 신경망의 출력인 O_{M_s}로부터, 그리고 K,V는 D-MPNN의 출력인 O_{G_s}로부터 생성될 수 있다. O_{M_s} 및 O_{G_s}로부터 Q,K,V 값이 생성되는 과정은 전술한 수학식 1을 참조한다.

[0050] 일 실시 예에 따르면, 화합물 표현형을 개선하기 위해 교차 어텐션 블록의 출력을 셀프-어텐션(self-attention) 블록의 입력으로 하여 셀프-어텐션 매커니즘을 추가적으로 적용시켜 화합물의 최종 표현형 $Comp_s$를 생성할 수 있다.

[0051] 다시 도 3에 대한 설명으로 돌아가서, 장치는 단백질 시퀀스 데이터를 기반으로 단백질 정보를 인코딩 할 수 있다(S320). S320 단계에서 단백질 정보를 인코딩하기 위한 과정을 이하 도 5를 참조하여 설명한다.

[0052] 도 5는 본 발명의 일 실시 예에 따른 단백질 정보를 인코딩하는 과정을 나타낸 도면이다.

[0053] 도 5를 참조하면, 단백질 데이터는 residue를 UniRep Vocabulary에서 부여한 고유 숫자로 표현할 수 있는 TAPE(Tasks Assessing Protein Embeddings) tokenizer를 이용해 전처리 되며(S510), 이후 one-hot 인코딩을 기반으로 1차원 데이터인 단백질 시퀀스 데이터가 생성될 수 있다(S520). 생성된 단백질 시퀀스 데이터는 1-D CNN(convolutional neural network)에 입력되며, 1D CNN의 출력으로 최종 단백질 Feature인 $Prot_t$가 생성될 수 있다(S530).

[0054] S310 단계 및 S320 단계를 통해 인코딩된 화합물 정보 및 단백질 정보는 교차 어텐션 블록에 입력되며, 교차 어텐션 매커니즘을 통해 결합될 수 있다(S330). S330 단계에서 사용되는 교차-어텐션 블록은 Q(Query), K(Key), V(Value)를 입력값으로 가지며, Q 는 인코딩된 화합물 정보인 $Comp_s$로부터, 그리고 K,V는 인코딩된 단백질 정보인 $Prot_t$로부터 생성될 수 있다. $Comp_s$ 및 $Prot_t$로부터 Q,K,V를 생성하는 과정은 전술한 수학식 3을 참조한다.

[0055] 마지막으로, 장치 S330 단계에서 사용되는 교차 어텐션 블록의 출력을 기반으로 화학물과 단백질의 상호 작용을 예측할 수 있다(S340). 일 예시로, 도 1에 도시된 바와 같이 S330 단계의 교차 어텐션 블록의 출력은 MLP 인공 신경망을 거쳐 Binding Score를 생성할 수 있으며, 해당 Binding Score를 기반으로 최종적인 화합물과 단백질 간 상호 작용을 예측할 수 있다.

[0056] 도 6는 본 발명의 일 실시 예에 따른 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 장치의 구성을 나타낸 도면이다.

[0057] 본 발명의 일 실시 예에 따른 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 장치(600)는 프로세서(610), 네트워크 인터페이스(620), 메모리(630), 스토리지(640) 및 이들을 연결하는 데이터 버스(650)를 포함할 수 있으며, 기타 본 발명의 목적을 달성함에 있어 요구되는 부가적인 구성들을 더 포함할 수 있음은 물론이라 할 것이다.

[0058] 프로세서(610)는 각 구성의 전반적인 동작을 제어한다. 프로세서(610)는 CPU(Central Processing Unit), MPU(Micro Processer Unit), MCU(Micro Controller Unit) 또는 본 발명이 속하는 기술 분야에서 널리 알려져 있는 형태의 프로세서 중 어느 하나일 수 있다. 아울러, 프로세서(610)는 본 발명의 일 실시 예에 따른 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법을 수행하기 위한 적어도 하나의 애플리케이션 또는 프로그램에 대한 연산을 수행할 수 있다.

[0059] 네트워크 인터페이스(620)는 본 발명의 일 실시 예에 따른 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 장치(600)의 유무선 인터넷 통신을 지원하며, 그 밖의 공지의 통신 방식을 지원할 수도 있다. 따라서 네트워크 인터페이스(620)는 그에 따른 통신 모듈을 포함하여 구성될 수 있다.

[0060] 메모리(630)는 각종 정보, 명령 및/또는 정보를 저장하며, 본 발명의 일 실시 예에 따른 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 방법을 수행하기 위해 스토리지(640)로부터 하나 이상의 컴퓨터 프로그램(641)을 로드할 수 있다. 도 1에서는 메모리(630)의 하나로 RAM을 도시하였으나 이와 더불어 다양한 저장 매체를 메모리(630)로 이용할 수 있음은 물론이다.

[0061] 스토리지(640)는 하나 이상의 컴퓨터 프로그램(641) 및 대용량 네트워크 정보를 비임시적으로 저장할 수 있다. 이러한 스토리지(640)는 ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 메모리 등과 같은 비휘발성 메모리, 하드 디스크(HDD), 보조 저장 매체(SSD), 착탈형 디스크, 또는 본 발명이 속하는 기술 분야에서 널리 알려져 있는 임의의 형태의 컴퓨터로 읽을 수 있는 기록 매체 중 어느 하나일 수 있다.

[0062] 컴퓨터 프로그램(641)은 메모리(630)에 로드되어, 하나 이상의 프로세서(610)에 의해, 분자 그래프 데이터 및 분자 지문 데이터를 기반으로 화합물 정보를 인코딩하는 동작, 단백질 시퀀스 데이터를 기반으로 단백질 정보를 인코딩하는 동작, 인코딩된 화합물 정보 및 단백질 정보를 교차 어텐션 블록에 입력하는 동작 및 교차 어텐션 블록의 출력을 기반으로 화합물과 단백질의 상호 작용을 예측하는 동작을 실행할 수 있다.

[0063] 데이터 버스(650)는 이상 설명한 프로세서(610), 네트워크 인터페이스(620), 메모리(630) 및 스토리지(640) 사이의 명령 및/또는 정보의 이동 경로가 된다.

[0064] 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.

부호의 설명

[0065] 600: 교차 어텐션 매커니즘 기반의 화합물-단백질 상호작용 예측 장치

610: 프로세서

620: 네트워크 인터페이스

630: 메모리

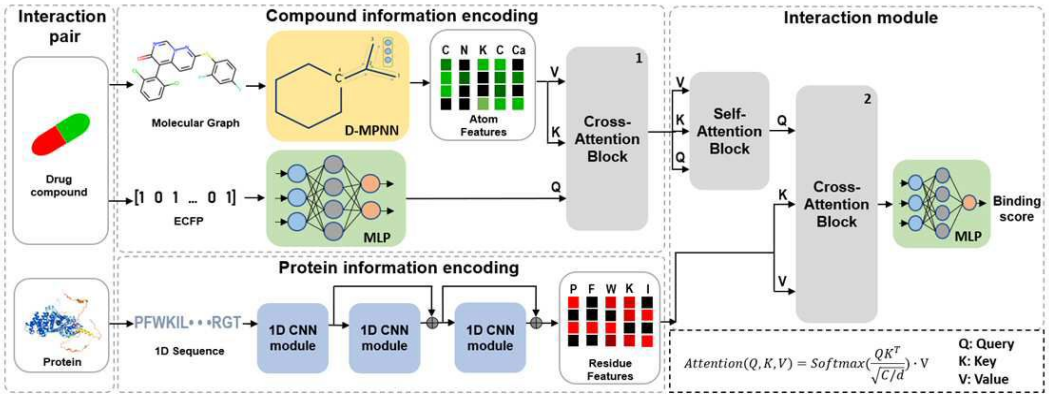
640: 스토리지

641: 컴퓨터 프로그램

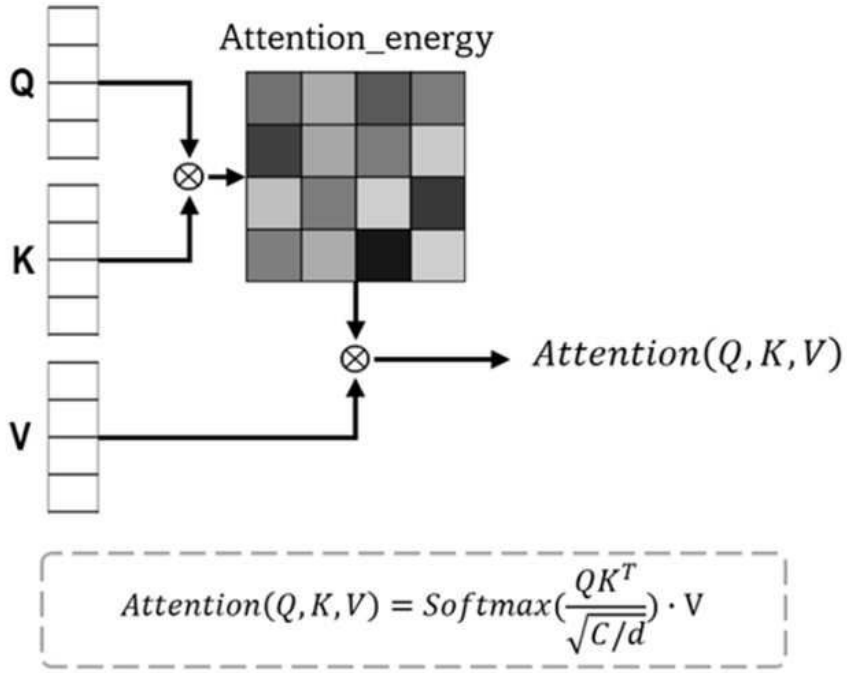
650: 데이터 버스

도면

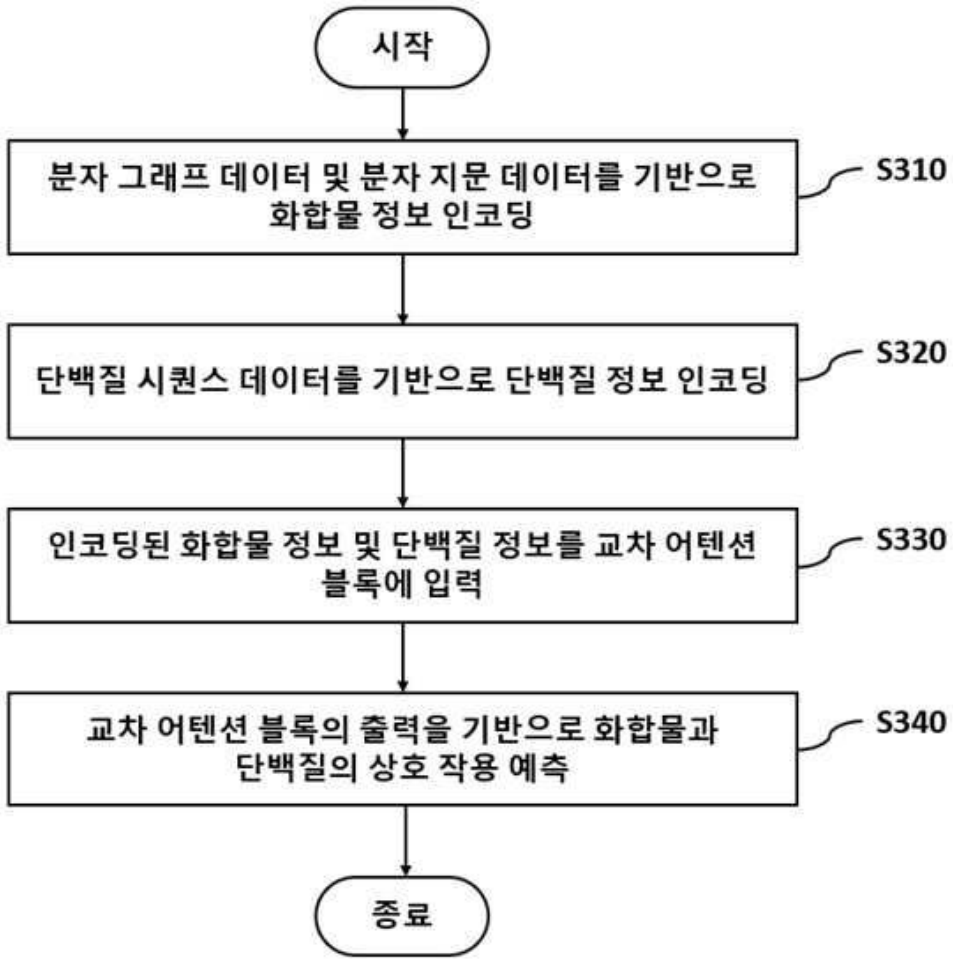
도면1



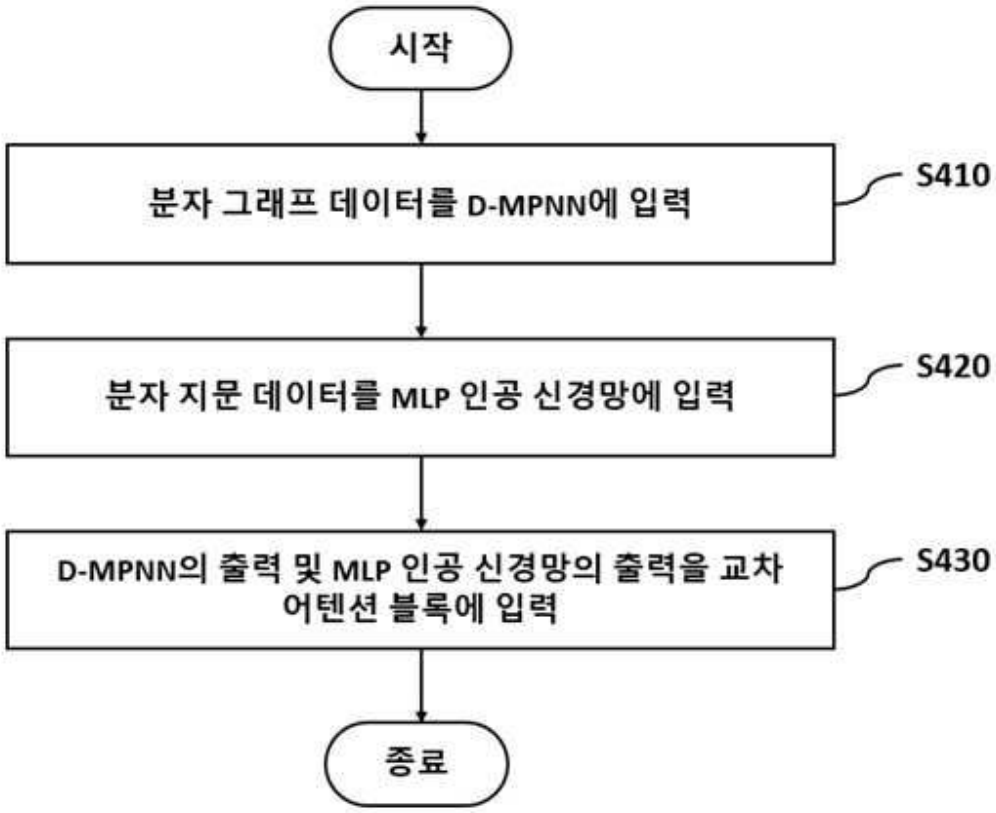
도면2


Attention_energy
Q
K
V
Attention(Q,K,V)
Attention(Q,K,V) = Softmax(QK^T / √(C/d)) · V

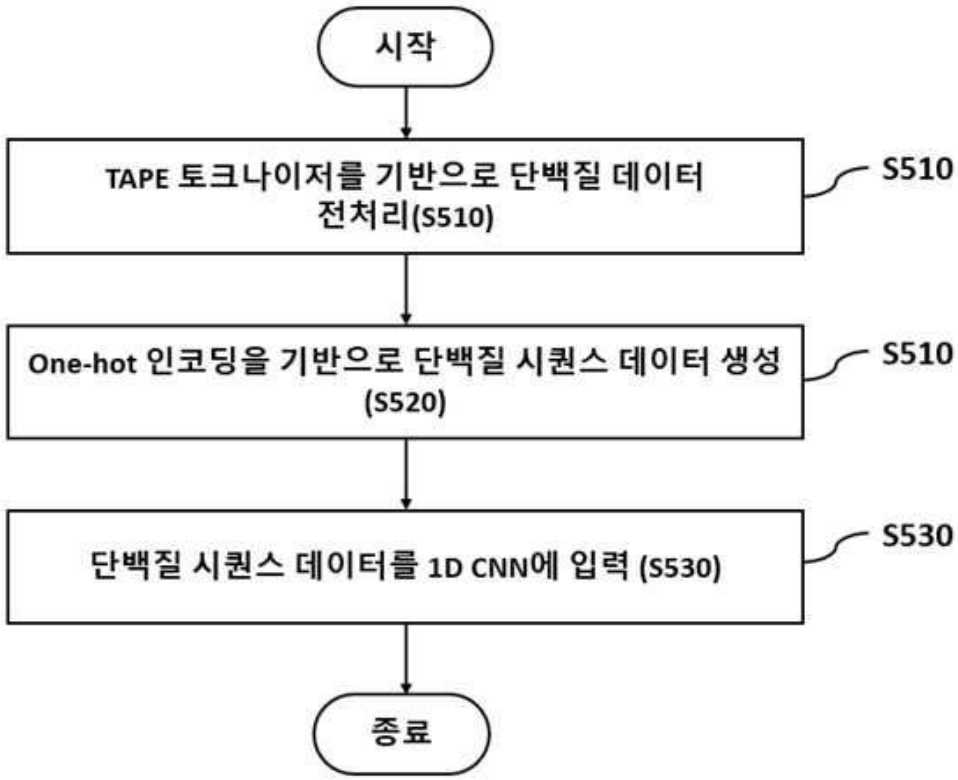
도면3


시작
분자 그래프 데이터 및 분자 지문 데이터를 기반으로 화합물 정보 인코딩
S310
단백질 시퀀스 데이터를 기반으로 단백질 정보 인코딩
S320
인코딩된 화합물 정보 및 단백질 정보를 교차 어텐션 블록에 입력
S330
교차 어텐션 블록의 출력을 기반으로 화합물과 단백질의 상호 작용 예측
S340
종료

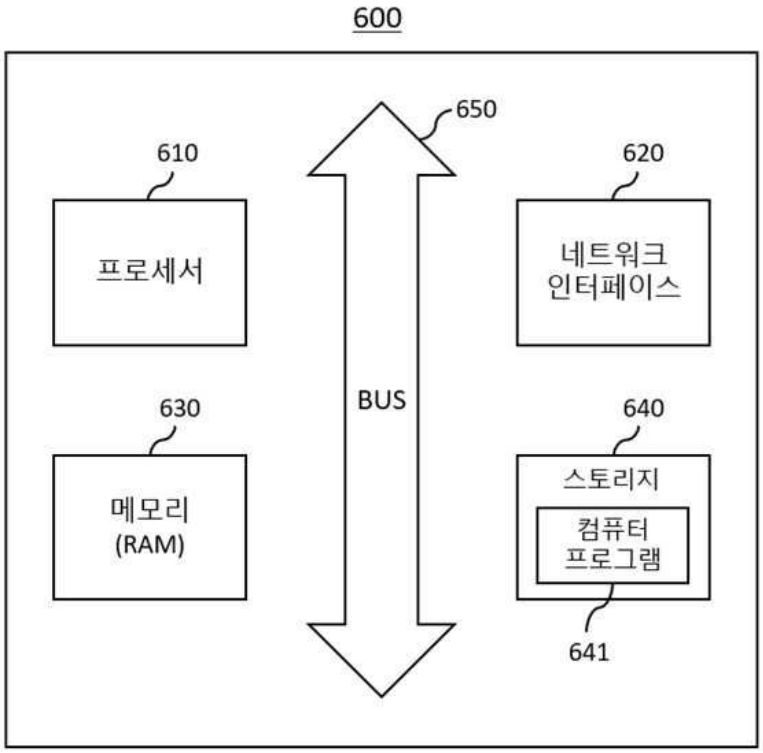
도면4


시작
분자 그래프 데이터를 D-MPNN에 입력
S410
분자 지문 데이터를 MLP 인공 신경망에 입력
S420
D-MPNN의 출력 및 MLP 인공 신경망의 출력을 교차 어텐션 블록에 입력
S430
종료

도면5


시작
TAPE 토크나이저를 기반으로 단백질 데이터 전처리(S510)
S510
One-hot 인코딩을 기반으로 단백질 시퀀스 데이터 생성 (S520)
S510
단백질 시퀀스 데이터를 1D CNN에 입력 (S530)
S530
종료

도면6


600
650
610
620
프로세서
네트워크
인터페이스
BUS
630
640
메모리
(RAM)
스토리지
컴퓨터
프로그램
641