

【서지사항】

【서류명】	특허출원서
【참조번호】	SDP20244683
【출원구분】	특허출원
【출원인】	
【명칭】	연세대학교 산학협력단
【특허고객번호】	2-2005-009509-9
【대리인】	
【명칭】	특허법인시공
【대리인번호】	9-2023-100041-2
【지정된변리사】	조예찬, 손하윤
【포괄위임등록번호】	2023-059479-9
【발명의 국문명칭】	화합물-단백질 결합 친화도 예측 방법 및 장치
【발명의 영문명칭】	METHOD FOR PREDICITING COMPOUND-PROTEIN BINDING AFFINITY AND DEVICE THEREOF
【발명자】	
【성명】	박상현
【성명의 영문표기】	SANGHYUN PARK
【국적】	KR
【주민등록번호】	670101-1XXXXXX
【우편번호】	08004
【주소】	서울특별시 양천구 오목로 300, 204동 3701호

【거주국】 KR

【발명자】

【성명】 서상민

【성명의 영문표기】 SANGMIN SEO

【국적】 KR

【주민등록번호】 930507-1XXXXXX

【우편번호】 63272

【주소】 제주특별자치도 제주시 고마로 44, 801호

【거주국】 KR

【발명자】

【성명】 최승연

【성명의 영문표기】 SEUNGYEON CHOI

【국적】 KR

【주민등록번호】 970828-1XXXXXX

【우편번호】 07213

【주소】 서울특별시 영등포구 당산로45길 7-3, 101동 808호

【거주국】 KR

【발명자】

【성명】 김환희

【성명의 영문표기】 HWANHEE KIM

【국적】 KR

【주민등록번호】 980813-2XXXXXX

【우편번호】 03788

【주소】 서울특별시 서대문구 연세로11길 26, 4층 401호

【거주국】 KR

【발명자】

【성명】 이지은

【성명의 영문표기】 JIEUN LEE

【국적】 KR

【주민등록번호】 960912-2XXXXXX

【우편번호】 03716

【주소】 서울특별시 서대문구 동교로 291, 101동 1004호

【거주국】 KR

【출원언어】 국어

【심사청구】 청구

【이 발명을 지원한 국가연구개발사업】

【과제고유번호】 1711198526

【과제번호】 00229822

【부처명】 과학기술정보통신부

【과제관리(전문)기관명】 한국연구재단

【연구사업명】 인공지능활용혁신신약발굴

【연구과제명】 난치성 질환 극복을 위한 인공지능 기반의 다중 약물 적응
증 최적화 플랫폼 개발 및 혁신신약 발굴

【과제수행기관명】 연세대학

【연구기간】 2024.01.01 ~ 2024.12.31

【취지】 위와 같이 특허청장에게 제출합니다.

대리인 특허법인시공 (서명 또는 인)

【수수료】

【출원료】 0 면 46,000 원

【가산출원료】 41 면 0 원

【우선권주장료】 0 건 0 원

【심사청구료】 10 항 676,000 원

【합계】 722,000원

【감면사유】 전담조직(50%감면)[1]

【감면후 수수료】 361,000 원

【발명의 설명】

【발명의 명칭】

화합물-단백질 결합 친화도 예측 방법 및 장치{METHOD FOR PREDICITING COMPOUND-PROTEIN BINDING AFFINITY AND DEVICE THEREOF}

【기술분야】

【0001】 본 발명은 화합물-단백질 결합 친화도 예측 방법 및 장치에 관한 것으로, 보다 구체적으로, 화합물-단백질 복합체 자유 기반으로 화합물-단백질 결합 친화도를 정확하고 효과적으로 예측하는 화합물-단백질 결합 친화도 예측 방법에 관한 것이다.

【발명의 배경이 되는 기술】

【0003】 약물 개발 초기 단계에서 고속 처리 스크리닝이 사용되지만, 모든 화합물-단백질 상호작용을 평가하는 것은 불가능하다. 이를 보완하기 위해 단백질 구조 기반 가상 스크리닝이 사용되며, 분자 도킹(molecular docking)이 널리 활용되지만 계산 자원이 많이 들고 정확성이 제한적이라는 한계가 있었다.

【0004】 또한, 실험적으로 얻어진 3D 화합물-단백질 복합체 구조의 부족으로 인해 훈련 데이터셋이 제한되어 있어 개발에 한계가 있었다.

【0005】 이에, 본 발명의 발명자들은 화합물-단백질 복합체 구조에 비의존적으로 화합물-단백질의 결합 친화도를 예측하고자 일응 노력한 결과 본 발명의 완성

하였다.

【발명의 내용】

【해결하고자 하는 과제】

【0007】 본 명세서에서 기재된 다양한 실시예들은 화합물-단백질 복합체 자유(complex free) 기반으로 화합물-단백질 결합 친화도를 정확하고 효과적으로 예측하는 화합물-단백질 결합 친화도 예측 방법을 제공하는 것을 목적으로 한다.

【0008】 본 개시가 해결하고자 하는 과제들은 상술한 과제에 제한되지 않으며, 언급되지 않은 또 다른 과제들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

【과제의 해결 수단】

【0010】 본 개시의 일 실시예에 따르면, 적어도 하나 이상의 프로세서에 의해 수행되는 화합물-단백질 결합 친화도 예측 방법은, 상호작용하는 화합물 데이터 및 단백질 데이터를 입력받는 단계; 상기 입력된 화합물 데이터 및 상기 단백질의 데이터를 기초로 화합물 속성 벡터 및 단백질 속성 벡터를 생성하는 단계; 상기 화합물의 속성 벡터와 상기 단백질의 속성 벡터를 기초로 어텐션 값을 산출하는 단계; 상기 어텐션 값을 기초로 제1 상호작용 행렬을 생성하는 단계; 상기 제1 상호작용 행렬을 학습 데이터로 하여 화합물-단백질의 비공유 상호작용(non-covalent

interaction) 및 결합 친화도(binding affinity)을 예측하도록 제1 인공지능 모델을 학습시키는 단계; 및 상기 제1 인공지능 모델의 출력값을 기초로 화합물-단백질의 비공유 상호작용 및 결합 친화도를 예측하는 단계;를 포함한다.

【0011】 본 개시의 일 실시예에 따르면, 상기 화합물 데이터를 입력받는 단계는 화합물 구조 그래프를 생성하는 단계를 포함하며, 상기 화합물 구조 그래프는 원자의 속성을 노드로, 결합(bond)의 속성을 엣지로 한 것이다.

【0012】 본 개시의 일 실시예에 따르면, 상기 화합물의 속성 벡터와 상기 단백질의 속성 벡터를 기초로 어텐션 값을 산출하는 단계는, 상기 화합물의 속성 벡터와 상기 단백질의 속성 벡터를 기초로 각각의 키(key), 벨류(value), 및 쿼리(query)를 산출하는 단계; 및 상기 산출된 화합물의 쿼리와 단백질의 쿼리를 교차하거나, 화합물의 키 및 벨류와 단백질의 키 및 벨류를 교차해 서브 어텐션 레이어에 제공하는 단계;를 포함한다.

【0013】 본 개시의 일 실시예에 따르면, 상기 화합물의 속성 벡터와 상기 단백질의 속성 벡터를 기초로 어텐션 값을 산출하는 단계는, 상기 서브 어텐션 레이어의 출력값을 셀프 어텐션 레이어에 제공하는 단계; 및 상기 셀프 어텐션 레이어의 출력값을 피드-포워드 레이어(feed-forward layer)에 제공하는 단계;를 더 포함한다.

【0014】 본 개시의 일 실시예에 따르면, 상기 제1 인공지능 모델을 학습시키는 단계는, 상기 화합물의 하나 이상의 원소 및 상기 단백질의 하나 이상의 잔기 사이의 상호작용 점수를 산출하는 단계; 상기 점수에서 잠재변수를 추출하는 단계;

및 상기 잠재변수를 완전 연결 레이어(fully connected layer)에 제공하는 단계;를 포함한다.

【0015】 본 개시의 일 실시예에 따르면, 상기 제1 인공지능 모델의 출력값은 제1 인공지능 모델에서 예측된 결합 친화도, 상기 화합물의 속성 벡터, 상기 단백질의 속성 벡터, 및 상기 제1 상호작용 행렬을 포함한다.

【0016】 본 개시의 일 실시예에 따르면, 상기 제1 인공지능 모델의 출력값을 기초로 화합물-단백질의 비공유 상호작용(non-covalent interaction) 및 결합 친화도를 예측하는 단계는, 상기 제1 인공지능 모델의 출력값을 제2 인공지능 모델에 입력해 지식 증류를 수행하는 단계; 및 상기 제2 인공지능 모델에서 비공유 상호작용(non-covalent interaction) 및 결합 친화도를 예측을 수행하는 단계를 포함하고, 상기 제1 인공지능 모델이 입력받은 데이터는 화합물-단백질 복합체(compound-protein complex)에 기초한 데이터인 것이다.

【0017】 본 개시의 일 실시예에 따르면, 상기 지식 증류를 수행하는 단계는, 제1 상호작용 행렬과 상기 제2 인공지능 모델에서 생성한 제2 상호작용 행렬을 비교하여 손실 함수를 산출하는 단계를 포함한다.

【0018】 본 개시의 일 실시예에 따르면, 상술된 방법을 컴퓨터에서 실행시키도록 판독 가능한 기록 매체에 저장된 컴퓨터 프로그램이 제공된다.

【0019】 본 개시의 일 실시예에 따르면, 컴퓨팅 장치로서, 통신 모듈; 메모리; 및 상기 메모리와 연결되고, 상기 메모리에 포함된 컴퓨터 판독 가능한 적어도

하나의 프로그램을 실행하도록 구성된 적어도 하나의 프로세서;를 포함하고, 상기 적어도 하나의 프로그램은, 상호작용하는 화합물 및 단백질의 구조를 입력받고, 상기 입력된 화합물 구조 및 상기 단백질의 구조를 기초로 화합물 속성 벡터 및 단백질 속성 벡터를 생성하고, 상기 화합물의 속성 벡터와 상기 단백질의 속성 벡터를 기초로 어텐션 값을 산출하고, 상기 어텐션 값을 기초로 제1 상호작용 행렬을 생성하고, 상기 제1 상호작용 행렬을 학습 데이터로 하여 화합물-단백질의 비공유 상호작용(non-covalent interaction) 및 결합 친화도(binding affinity)을 예측하도록 제1 인공지능 모델을 학습시키고, 상기 제1 인공지능 모델의 출력값을 기초로 화합물-단백질의 비공유 상호작용 및 결합 친화도를 예측하는 컴퓨팅 장치를 제공한다.

【발명의 효과】

【0021】 본 개시의 일 실시예에 따르면, 화합물-단백질 복합체 자유 기반으로 화합물-단백질 결합 친화도를 정확하고 효과적으로 예측할 수 있다.

【0022】 본 개시에 따른 효과들은 상술한 효과에 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

【도면의 간단한 설명】

【0024】 도 1은 일 실시예에 따른 컴퓨팅 장치의 내부 구성을 나타내는 기능적인 블록도이다.

도 2는 일 실시예에 따른 인공지능 모델의 내부 구성을 나타내는 기능적인 블록도이다.

도 3은 일 실시예에 따른 화합물-단백질 결합 친화도 예측 방법의 순서도이다.

도 4는 일 실시예에 따른 레이어를 이용하여 어텐션 값을 산출하는 단계를 나타내는 도면이다.

도 5는 일 실시예에 따른 화합물-단백질의 결합 친화도를 예측하는 모델들 간의 성능을 비교한 그래프이다.

도 6은 일 실시예에 따른 컴퓨팅 장치의 하드웨어 구성을 나타내는 블록도이다.

【발명을 실시하기 위한 구체적인 내용】

【0025】 이하, 첨부된 도면들에 기재된 내용들을 참조하여 본 발명에 따른 예시적 실시 예를 상세하게 설명한다. 다만, 본 발명이 예시적 실시 예들에 의해 제한되거나 한정되는 것은 아니다. 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 개시가 속하는 기술 분야에서 통상의 지식을 가진 자에게 공통적으로 이해될 수 있는 의미로 사용될 것이나, 이는 당 분야에 종사하는 기술자의 의도 또는 관례, 새로운 기술의 출현 등에 따라 달라질 수

있다.

【0026】 또한, 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다. 특정한 경우, 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서, 본 개시에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 개시의 전반에 걸친 내용을 토대로 정의되어야 한다.

【0027】 본 명세서 전체에서 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라, 다른 구성요소를 더 포함할 수 있음을 의미한다. 또한, 본 명세서에서 사용된 단수형은 특별히 언급하지 않는 한 복수형도 포함한다. 또한, 본 명세서 전체에서 기재된 "a, b 및/또는 c 중 적어도 하나"의 표현은, 'a 단독', 'b 단독', 'c 단독', 'a 및 b', 'a 및 c', 'b 및 c', 또는 'a, b, c 모두'를 포괄할 수 있다.

【0028】 한편, 본 명세서에서 사용되는 "제1 및/또는 제2" 등의 용어는 다양한 구성요소들을 설명하기 위하여 사용될 수 있으나, 이는 하나의 구성요소를 다른 구성요소로부터 구별하기 위한 목적으로만 사용될 뿐, 해당 용어로 지칭되는 구성요소로 한정하기 위한 것은 아니다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않는 한, 제1 구성요소는 제2 구성요소로 명명될 수 있으며, 제2 구성요소 또한 제1 구성요소로 명명될 수 있다.

【0029】 또한, 본 명세서에 기재된 "...부", "...모듈" 등의 용어는 적어

도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어 또는 소프트웨어로 구현되거나 하드웨어와 소프트웨어의 결합으로 구현될 수 있다. 또한, 본 명세서에서 본 개시의 실시 예는 기능적인 블록 구성들 및 다양한 처리 단계들로 나타내어질 수 있다. 이러한 기능 블록들은 특정 기능들을 실행하는 다양한 개수의 하드웨어 또는/및 소프트웨어 구성들로 구현될 수 있다. 예를 들어, 본 개시의 실시 예는 하나 이상의 마이크로프로세서의 제어 또는 다른 제어 장치들에 의해서 다양한 기능들을 실행할 수 있는, 메모리, 프로세싱, 로직(logic), 룩 업 테이블(look-up table) 등과 같은 직접 회로 구성들을 채용할 수 있다.

【0030】 본 개시에 따른 실시 예에서, 인공지능과 관련된 기능은 프로세서 및 메모리를 통해 구현될 수 있다. 이 때, 프로세서는 CPU(Center Processing Unit), AP(Application Processor), DSP(Digital Signal Processor) 등과 같은 범용 프로세서, GPU(Graphic Processing Unit), VPU(Vision Processing Unit)와 같은 그래픽 전용 프로세서 및 NPU(Neural network Processing Unit)와 같은 인공지능 전용 프로세서 중 어느 하나일 수 있다. 프로세서는 메모리에 저장된 기 정의된 동작 규칙 또는 인공지능 모델에 따라 입력 데이터를 처리할 수 있다. 또는, 프로세서가 인공지능 전용 프로세서인 경우, 인공지능 전용 프로세서는 특정 인공지능 모델의 처리에 특화된 하드웨어 구조로 설계될 수 있다. 본 개시에 따른 일부 실시 예에서, 인공지능과 관련된 기능은 복수의 프로세서들을 통해 구현될 수 있다.

【0031】 본 개시에 따른 실시 예에서, 기 정의된 동작 규칙 또는 인공지능 모델은 기계학습을 수행하도록 구성될 수 있다. 여기서, 기계학습을 수행하도록 구

성된다는 것은, 기 정의된 동작 규칙 또는 인공지능 모델이 학습 알고리즘을 기반으로 다수의 학습 데이터들을 이용하여 학습되어 원하는 특성(또는 목적)을 수행하도록 구성됨을 의미한다. 이러한 학습은 본 개시에 따른 인공지능이 구현되는 장치 자체에서 이루어질 수도 있고, 별도의 서버 및/또는 시스템을 통해 이루어질 수도 있다.

【0032】 인공지능 모델은 뉴럴 네트워크(또는 인공 신경망)로 구현될 수 있으며, 기계학습과 인지과학에서 생물학의 신경을 모방한 통계학적 학습 알고리즘에 기반하여 동작할 수 있다. 뉴럴 네트워크는 시냅스의 결합으로 네트워크를 형성한 인공 뉴런(노드)이 학습을 통해 시냅스의 결합 세기를 변화시켜 문제 해결 능력을 가지는 모델 전반을 의미할 수 있다. 뉴럴 네트워크는 복수의 신경망 레이어(layer)들로 구성될 수 있으며, 예시적으로 뉴럴 네트워크는 입력 레이어(input layer), 은닉 레이어(hidden layer) 및 출력 레이어(output layer)를 포함할 수 있다. 복수의 신경망 레이어들 각각은 적어도 하나의 노드(node) 및 적어도 하나의 가중치(weight)를 포함할 수 있으며, 이전(precious) 레이어의 연산 결과와 가중치 간의 연산을 통해 신경망 연산을 수행할 수 있다. 복수의 신경망 레이어들이 가지고 있는 적어도 하나의 가중치는 인공지능 모델의 학습 결과에 의하여 최적화될 수 있다. 예를 들어, 학습 과정동안 인공지능 모델에서 획득한 손실(loss) 값 또는 비용(cost) 값이 감소 또는 최소화되도록 적어도 하나의 가중치가 갱신될 수 있다. 뉴럴 네트워크는 임의의 입력으로부터 예측하고자 하는 결과를 추론할 수 있다.

【0033】 인공지능 모델의 학습 방법은 학습 방식에 따라 입력 데이터 및 출력 데이터가 훈련 데이터로써 제공되어 문제(입력 데이터)에 대응하는 정답(출력 데이터)이 정해져 있는 지도학습(supervised learning), 출력 데이터 없이 입력 데이터만 제공되어 문제(입력 데이터)에 대응하는 정답(출력 데이터)이 정해지지 않은 비지도학습(unsupervised learning) 및 현재 상태(state)에서 어떤 행동(action)을 취할 때마다 보상(reward)이 부여되고, 이러한 보상을 최대화하는 방향으로 학습을 진행하는 강화학습(reinforcement learning) 등으로 구분될 수 있다. 또는, 학습 모델의 구조인 아키텍처에 따라 구분될 수도 있다.

【0034】 본 개시의 실시 예에서, 인공지능 모델은 GoogleNet, AlexNet, VGG Network 등과 같은 CNN(Convolution Neural Network), R-CNN(Region with Convolution Neural Network), RPN(Region Proposal Network), RNN(Recurrent Neural Network), S-DNN(Stacking-based deep Neural Network), S-SDNN(State-Space Dynamic Neural Network), Deconvolution Network, DBN(Deep Belief Network), RBM(Restrcted Boltzman Machine), Fully Convolutional Network, LSTM(Long Short-Term Memory) Network, Classification Network, Generative Modeling, eXplainable AI, Continual AI, Representation Learning, AI for Material Design, 자연어 처리를 위한 BERT, SP-BERT, MRC/QA, Text Analysis, Dialog System, GPT-3, GPT-4, 비전 처리를 위한 Visual Analytics, Visual Understanding, Video Synthesis, ResNet 데이터 지능을 위한 Anomaly Detection, Prediction, Time-Series Forecasting, Optimization, Recommendation, Data

Creation 등 다양한 인공지능 구조 및 알고리즘 중 적어도 하나를 이용할 수 있으며, 상술한 예시들은 본 개시의 실시 예에 따라 이용되는 인공지능 구조 및 알고리즘의 예를 나열한 것인 뿐, 본 개시의 실시 예에 따라 이용되는 인공지능 구조 및 알고리즘을 제한하는 것은 아니다.

【0035】 이하, 본 개시의 다양한 실시 예들은 첨부된 도면을 참조하여 상세하게 설명될 것이다. 실시 예를 설명함에 있어서 본 발명이 속하는 기술 분야에 익히 알려져 있고 본 발명과 직접적으로 관련이 없는 기술 내용에 대해서는 설명을 생략할 것이다. 이는 불필요한 설명을 생략함으로써 본 발명의 요지를 흐리지 않고 더욱 명확히 전달하기 위함이다. 마찬가지로 이유로 첨부 도면에 있어서 일부 구성요소는 과장되거나 생략되거나 개략적으로 도시되었다. 또한, 각 구성요소의 크기는 실제 크기를 전적으로 반영하는 것이 아니다. 본 명세서에서, 전문에 걸쳐 동일한 참조 부호는 동일한 또는 대응하는 구성 요소를 지칭할 수 있다.

【0036】 본 명세서에서 용어, '상호작용(interaction)'은 화합물-단백질 간의 비공유 상호작용(non-covalent interaction)를, 구체적으로 화합물의 원소 및 단백질의 잔기 사이의 비공유 상호작용을 의미할 수 있다.

【0037】 본 명세서에서 용어, '결합 친화도(binding affinity)'는 화합물-단백질 간의 결합 상호작용의 세기를 의미할 수 있다.

【0038】 본 명세서에서 용어, '지식 증류(knowledge distillation)'은 교사 모델(teacher model)이 학생 모델(student model)로 학습한 지식을 전달하는 방법을 의미할 수 있다. 본 명세서에서 교사 모델은 제1 인공지능 모델(100), 학생 모

텔은 제2 인공지능 모델일 수 있다.

【0039】 도 1은 일 실시예에 따른 컴퓨팅 장치(1000)의 내부 구성을 나타내는 기능적인 블록도이다. 도 1을 참조하면, 컴퓨팅 장치(1000)는 단백질 데이터(101) 및 화합물 데이터(102)를 입력받아 해당 단백질-화합물의 비공유 결합성 및 결합 친화도를 예측하기 위한 임의의 인공지능 장치를 지칭할 수 있다. 컴퓨팅 장치(1000)는 제1 인공지능 모델(100) 및 제2 인공지능 모델(200)을 포함하며, 제1 인공지능 모델(100) 및 제2 인공지능 모델(200)은 단백질 데이터(101) 및 화합물 데이터(102)를 입력받을 수 있다.

【0040】 제1 인공지능 모델(100) 및 제2 인공지능 모델(200)은 물리적으로 분리된 구성인 것으로 언급되나, 이는 일부 실시예에 불과하며, 또 다른 실시예에 따르면 제1 인공지능 모델(100) 및 제2 인공지능 모델(200)은 논리적으로 분리된 구조일 수 있고, 이 경우, 하나의 서버에서 분리된 기능에 의하여 구현될 수 있다.

【0041】 제1 인공지능 모델(100)은 단백질 데이터 및 화합물 데이터를 이용하여 화합물-단백질의 비공유 상호작용 및 결합 친화도를 예측하도록 학습시킨 장치일 수 있다. 제1 인공지능 모델(100)은 출력값을 제2 인공지능 모델(200)에 전달해 지식 증류를 수행할 수 있다.

【0042】 제2 인공지능 모델(200)은 단백질 데이터, 화합물 데이터 및 제1 인공지능 모델(100)로부터 전달받은 데이터에 기초하여 화합물-단백질의 비공유 상호작용 및 결합 친화도를 예측할 수 있다.

【0043】 제1 인공지능 모델(100)의 제1 출력값(긴과선-점선-점선, 1) 및 제2 인공지능 모델(200)의 제1 출력값(긴과선-점선-점선, 2)을 기초로 손실 함수를 산출할 수 있다. 또한, 제1 인공지능 모델(100)의 제2 출력값(과선-점선, 3), 및 제2 인공지능 모델(200)의 제2 출력값(과선-점선, 4)을 기초로 비공유 상호작용 및 결합 친화도를 예측할 수 있다. 컴퓨팅 장치(1000)의 제1 인공지능 모델(100) 및 제2 인공지능 모델(200)를 이용해 화합물-단백질의 비공유 상호작용 및 결합 친화도를 예측하는 방법은 하기에서 상술하겠다.

【0045】 도 2는 일 실시예에 따른 인공지능 모델의 내부 구성을 나타내는 기능적인 블록도이다.

【0046】 일 실시예에 따르면, 제1 인공지능 모델(100) 및 제2 인공지능 모델(200)은 모두 단백질 인코딩 모듈(110), 화합물 인코딩 모듈(120), 어텐션 모듈(130) 및 예측 모듈(140) 등을 포함할 수 있다. 이하, 단백질 인코딩 모듈(110), 화합물 인코딩 모듈(120), 어텐션 모듈(130) 및 예측 모듈(140)의 설명은 제1 인공지능 모델(100), 및 제2 인공지능 모델(200) 모두에 해당한다.

【0047】 제1 인공지능 모델(100) 및 제2 인공지능 모델(200)은 단백질 데이터(101) 및 화합물 데이터(102)를 입력받거나 수신할 수 있다.

【0048】 제1 인공지능 모델(100)는 단백질 데이터(101) 및 화합물 데이터(102)를 화합물-단백질 복합체(compound-protein complex) 3차원 구조로부터 획득

할 수 있으며, 예를 들어, PDBbind database에서 획득할 수 있다.

【0049】 제2 인공지능 모델(200)은 단백질 데이터(101) 및 화합물 데이터(102)를 화합물-단백질 복합체(compound-protein complex) 자유(free) 모델로부터 획득할 수 있으며, 예를 들어, BindingDB database에서 획득할 수 있다.

【0050】 여기서, 단백질 데이터(101)는 단백질 표면의 움푹 파인 영역인 포켓(pocket) 영역의 리간드를 포함할 수 있고, 화합물 데이터(102)는 단백질과 결합하는 영역인 표적 분자 영역을 포함할 수 있다.

【0051】 단백질 데이터(101) 및 화합물 데이터(102)를 입력받는 경우, 단백질 인코딩 모듈(110) 및 화합물 인코딩 모듈(120)은 해당 단백질 데이터(101) 및 화합물 데이터(102)를 어텐션 모듈(130) 등의 입력으로 사용하기 위해 인코딩(encoding)을 수행할 수 있다. 여기서, 인코딩은 원래의 데이터를 다른 형식으로 변환하는 방법을 의미할 수 있다.

【0052】 단백질 인코딩 모듈(110) 및 화합물 인코딩 모듈(120)은 각각 단백질 데이터(101) 및 화합물 데이터(102)에 기초로 단백질 속성 벡터 및 화합물 속성 벡터를 생성해 어텐션 모듈(130)에 전송할 수 있다. 구체적으로, 단백질 인코딩 모듈(110) 및 화합물 인코딩 모듈(120)은 단백질 데이터(101) 및 화합물 데이터(102)를 기초로 임베딩(embedding)해 단백질 포켓의 잔기(residue) 정보 및 화합물의 원소 정보를 획득할 수 있다.

【0053】 여기서, 임베딩은 데이터를 수치화하기 위해 시퀀스(예를 들어, 아미노산 서열 또는 화합물의 원소)를 특정 수치 값을 갖는 벡터로 변환하는 것을 나타낼 수 있다 의미할 수 있다.

【0055】 단백질 속성 벡터 및 화합물 속성 벡터를 입력받는 경우, 어텐션 모듈(130)은 단백질 속성 벡터 및 화합물 속성 벡터를 기초로 어텐션 값을 산출할 수 있다. 여기서, 어텐션 값은 화합물 및 단백질의 상호작용을 나타낼 수 있다. 어텐션 모듈(130)은 단백질 속성 벡터를 입력받는 경우 단백질 어텐션 값을 산출해 예측 모듈(140)에 전송할 수 있다. 어텐션 모듈(130)은 화합물 속성 벡터를 입력받는 경우 화합물 어텐션 값을 산출해 예측 모듈(140)에 전송할 수 있다.

【0056】 단백질 어텐션 값 및 화합물 어텐션 값을 입력받는 경우, 예측 모듈(140)은 단백질 어텐션 값 및 화합물 어텐션 값에 기초하여 상호작용 행렬을 생성할 수 있다. 구체적으로, 제1 인공지능 모델(100)의 예측 모듈(140)은 제1 상호작용 행렬을, 제2 인공지능 모델(200)의 예측 모듈(140)은 제2 상호작용 행렬을 생성할 수 있다. 여기서, 상호작용 행렬은 단백질의 잔기와 화합물의 원소와의 상호작용 점수를 나타내는 행렬일 수 있다.

【0057】 도 1 및 도 2를 참조하면, 제1 인공지능 모델(100)은 제1 상호작용 행렬을 학습 데이터로 하여 화합물-단백질의 비공유 상호작용(non-covalent interaction) 및 결합 친화도(binding affinity)을 예측하도록 학습할 수 있다. 제

1 인공지능 모델(100)은 화합물-단백질 복합체를 입력 받아 대응하는 화합물-단백질의 비공유 상호작용 예측값 및 결합 친화도 예측값을 제2 인공지능 모델(200)에 전송할 수 있다.

【0058】 제2 인공지능 모델(200)은 제1 인공지능 모델의 출력값 및 제2 상호작용 행렬을 기초로 화합물-단백질의 비공유 상호작용 및 결합 친화도를 학습할 수 있다.

【0059】 이때, 제2 인공지능 모델(200)은 복합체 자유 기반이지만 제1 인공지능 모델(100)로부터 복합체 기반 데이터를 지식 증류받아 화합물-단백질 결합 친화도를 정확하게 예측할 수 있다.

【0060】 일 실시예에 따른 화합물-단백질 결합 친화도 예측 방법의 구체적인 내용은 하기 도 3 및 도 4에서 상술하겠다.

【0062】 도 3은 일 실시예에 따른 화합물-단백질 결합 친화도 예측 방법의 순서도이다. 도 3을 참조하면, 결합 친화도 예측 방법은 화합물 및 단백질 데이터를 입력하는 단계(S100), 화합물 속성 벡터 및 단백질 속성 벡터를 생성하는 단계(S200), 벡터를 기초로 어텐션 값을 산출하는 단계(S300), 제1 상호작용 행렬을 생성하고 제1 인공지능 모델을 학습하는 단계(S400), 제2 상호작용 행렬을 생성하고 제2 인공지능 모델을 학습하는 단계(S500), 및 제2 인공지능 모델에서 예측하는 단계(S600)를 포함할 수 있다. 이하 S100 단계 내지 S300 단계는 제1 인공지능 모델

(100) 및 제2 인공지능 모델(200) 모두에서 수행될 수 있다.

【0063】 S100 단계는 인공지능 모델(100 및 200)의 인코딩 모듈(110, 및 120)이 화합물 데이터 및 단백질 데이터를 입력받는 단계일 수 있다. 여기서 화합물 데이터 및 단백질 데이터는 제1 인공지능 모델(100)에서는 화합물-단백질 복합체 데이터로 제공될 수 있고, 제2 인공지능 모델(200)에서는 화합물-단백질 복합체 자유 데이터로 제공될 수 있다.

【0064】 입력받는 단백질 데이터는 화합물과 결합하는 포켓 영역의 잔기 아미노산 서열을 포함할 수 있다. 여기서, 제2 인공지능 모델(200)은 화합물-단백질 복합체 자유 데이터를 입력받아 입력된 단백질의 결합 포켓과 관련된 아미노산 서열을 추출하기 위해 단백질 결합 포켓 예측 도구를 사용할 수 있다. 이후, 제2 인공지능 모델(200)은 기 학습된 다른 인공지능 모델을 사용하여 포켓 영역의 잔기 아미노산 임베딩을 추출할 수 있다.

【0065】 일 실시예에서, 화합물 데이터를 입력받는 단계는 화합물 그래프를 생성하는 단계를 포함할 수 있다. 화합물 그래프는 공지된 인공지능 그래프를 사용하여 생성할 수 있으며, 여기서, 화합물 구조 그래프는 원자의 속성을 노드로, 결합(bond)의 속성을 엣지로 할 수 있다.

【0066】 이때, 화합물 데이터는 원소 속성 및 결합 속성을 포함할 수 있다. 예를 들어, 원소 속성은 원소의 유형, 키랄성, 형식 전하(formal charge), 결합된 수소 원소 수, 결합된 자유 전자의 수, 혼성화(hybridization), 방향족성(aromaticity), 고리성(cyclicality) 등을 포함할 수 있다. 여기서, 결합(bond)은 화

합물의 원소들 간의 결합으로 예를 들어, 결합(bond) 속성은 결합의 유형(단일 결합, 이중 결합, 삼중 결합), 결합의 특성(방향성, 키랄성, 스테레오, 콘쥬게이션) 등을 포함할 수 있다.

【0068】 S200 단계는 인공지능 모델(100 및 200)의 인코딩 모듈(110, 및 120)에서 화합물 속성 벡터 및 단백질 속성 벡터를 생성하는 단계일 수 있다.

【0069】 구체적으로, 단백질 인코딩 모듈(110)에서 잔기의 아미노산 임베딩을 입력으로 단백질 속성 벡터를 추출할 수 있다. 예를 들어, 잔기의 아미노산 임베딩을 입력으로 단백질 속성 벡터를 하기 수학식 1을 기초로 추출될 수 있다.

【0070】 【수학식 1】

$$p_j = SelfAttLayer_{P \rightarrow P} \left(x_j, \{x_1, \dots, x_{L_p}\} \right)$$

【0071】 여기서, p_j 는 추출된 단백질 속성 벡터, x_j 는 잔기의 아미노산을 나타낼 수 있다.

【0072】 화합물 인코딩 모듈(120)에서 화합물 구조 그래프의 노드 및 엣지를 입력으로 화합물 속성 벡터를 추출할 수 있다. 예를 들어, 화합물 구조 그래프의 노드 및 엣지를 입력으로 화합물 속성 벡터 하기 수학식 2를 기초로 추출될 수 있다.

【0073】 【수학식 2】

$$h_v^{(l)} = \text{COMBINE}(h_v^{(l-1)}, \text{AGGREGATE}(\{(h_v^{(l-1)}, h_u^{(l-1)}, e_{uv}) : u \in \mathcal{N}(v)\}))$$

【0074】 여기서, $\mathcal{N}(v)$ 은 노드 v 의 모든 이웃, e_{uv} 는 노드 u 와 v 의 엣지, $h_v^{(l-1)}$ 은 노드 v 의 $(l-1)$ 번째 레이어를 나타낼 수 있다.

【0075】 이 레이어를 반복적으로 적용하여, 노드를 업데이트 할 수 있고, 업데이트된 노드는 화합물의 속성 벡터로 사용될 수 있다.

【0077】 S300 단계는 인공지능 모델(100 및 200)의 어텐션 모듈(130)에서 화합물 속성 벡터 및 단백질 속성 벡터를 기초로 어텐션 값을 산출하는 단계일 수 있다.

【0078】 구체적으로, 어텐션 모듈(130)은 화합물의 속성 벡터와 상기 단백질의 속성 벡터를 기초로 각각의 키(key), 벨류(value), 및 쿼리(query)를 산출할 수 있다. 여기서, 각각의 키, 벨류, 및 쿼리는 공지된 어텐션 알고리즘을 사용하여 산출할 수 있다.

【0079】 어텐션 모듈(130)은 교차-어텐션(cross-attention) 레이어를 포함하여 교차-어텐션을 수행하여, 산출된 화합물의 쿼리와 단백질의 쿼리를 교차하거나, 화합물의 키 및 벨류와 단백질의 키 및 벨류를 교차해 서브 어텐션(sub attention)

레이어에 제공할 수 있다.

【0080】 서브 어텐션(sub attention) 레이어는 멀티 헤드 어텐션(multi-head self-attention) 레이어로 어텐션을 수행해 출력값을 셀프 어텐션(self attention) 레이어에 제공할 수 있다. 셀프 어텐션 레이어는 셀프 어텐션을 수행해 출력값을 피드-포워드 레이어(feed-forward)에 제공할 수 있다. 피드-포워드 레이어는 피드-포워드를 수행해 화합물 어텐션 값 및 단백질 어텐션 값을 제공할 수 있다. 따라서, 어텐션 모듈(130)이 화합물 속성 벡터를 입력받은 경우 출력값은 화합물 어텐션 값, 단백질 속성 벡터를 입력받은 경우 출력값은 단백질 어텐션 값일 수 있다.

【0081】 도 4는 일 실시예에 따른 레이어를 이용하여 어텐션 값을 산출하는 단계를 나타내는 도면이다. 도 4를 참조하면, 어텐션 모듈(130)은 화합물 속성 벡터(v)의 쿼리와 단백질 속성 벡터(p)의 쿼리를 교차하거나, 화합물 속성 벡터(v)의 키 및 벨류와 단백질 속성 벡터(p)의 키 및 벨류를 교차할 수 있다. 어텐션 모듈(130)은 서브 어텐션 레이어, 셀프 어텐션 레이어, 및 피드 포워드 레이어를 포함하여, 멀티 헤드 어텐션, 셀프 어텐션 및 피드 포워드를 사용하여 화합물 어텐션 값 및 단백질 어텐션 값을 산출할 수 있다. 최종 산출된 화합물 어텐션 값은 \bar{v}_i 로, 단백질 어텐션 값은 \bar{p}_j 로 나타낼 수 있다.

【0083】 S400 단계는 제1 인공지능 모델(100)의 예측 모듈(140)이 제1 상호작용 행렬을 생성하고 제1 인공지능 모델을 학습시키는 단계일 수 있다. 여기서, 제1 상호작용 행렬은 화합물의 원소 및 단백질의 잔기 사이의 상호작용 점수를 나타내는 것일 수 있다. 상호작용의 점수는 하기 수학식 3으로 산출할 수 있다.

【0084】 【수학식 3】

$$M_{ij} = \sigma(\bar{v}_i W_1) \cdot \sigma(\bar{p}_j W_2)^T$$

【0085】 여기서, M_{ij} 는 i번째 원소와 j번째 잔기 사이의 상호작용 점수, W_1 , 및 W_2 는 각각 학습 파라미터를 나타낼 수 있다.

【0086】 제1 인공지능 모델(100)의 예측 모듈(140)은 산출한 상호작용 점수에 기초하여 잠재변수를 추출할 수 있다.

【0087】 【수학식 4】

$$V_{inter} = \sum_{i=1}^{N_a} \sum_{j=1}^{L_p} M_{ij} [\sigma(\bar{v}_i W_3), \sigma(\bar{p}_j W_4)]$$

【0088】 여기서, V_{inter} 는 잠재변수, W_3 , 및 W_4 는 각각 학습 파라미터를 나타낼 수 있다.

【0089】 제1 인공지능 모델(100)의 예측 모듈(140)은 잠재변수를 완전 연결 레이어(fully connected layer)에 제공하여 결합 친화도를 예측할 수 있다.

【0090】 여기서, 제1 인공지능 모델(100)의 손실함수는 하기 수학식 5로 나타낼 수 있다.

【0091】 【수학식 5】

$$\mathcal{L}_{total} = \mathcal{L}_{aff} + \alpha_T \mathcal{L}_{inter}$$

【0092】 여기서, α_T 는 가중치 파라미터를 나타내고, \mathcal{L}_{aff} 및 \mathcal{L}_{inter} 는 각각 수학식 6 및 7으로 나타낼 수 있다.

【0093】 【수학식 6】

$$\mathcal{L}_{aff} = \frac{1}{N} \sum_{n=1}^N \|P_{gt} - P_{pred}\|_n^2$$

【0094】 【수학식 7】

$$\mathcal{L}_{inter} = \frac{1}{N} \sum_{n=1}^N \sum_{i=1}^{N_a} \sum_{j=1}^{L_p} -\|B_{ij} \log M_{ij} + (1 - B_{ij}) \log (1 - M_{ij})\|_n$$

【0095】 여기서, P_{gt} 는 실제 결합 친화도를, P_{pred} 는 예측된 결합 친화도를, B_{ij} 은 실제 i 번째 원소와 j 번째 잔기 사이의 바이너리 레이블을 나타낼 수 있다.

【0096】 제1 인공지능 모델(100)은 제1 인공지능 모델의 출력값을 제2 인공지능 모델(200)로 전달할 수 있다. 여기서, 제1 인공지능 모델의 출력값은 예측된 결합 친화도, 화합물의 속성 벡터, 단백질의 속성 벡터, 및 제1 상호작용 행렬을 포함할 수 있다.

【0098】 S500 단계는 제2 인공지능 모델(200)의 예측 모듈(140)이 제2 상호작용 행렬을 생성하고 제2 인공지능 모델을 학습시키는 단계일 수 있다. 구체적으로, 제1 인공지능 모델의 출력값을 제2 인공지능 모델에 입력해 지식 증류를 수행하는 단계로, 제2 인공지능 모델(200)의 예측 모듈(140)이 제2 상호작용 행렬을 생성하고 제1 인공지능 모델의 출력값에 기초하여 제2 인공지능 모델을 학습하는 단계일 수 있다. 여기서, 제1 인공지능 모델의 출력값은 예측된 결합 친화도, 화합물의 속성 벡터, 단백질의 속성 벡터, 및 제1 상호작용 행렬을 포함할 수 있다.

【0099】 제2 인공지능 모델(200)의 예측 모듈(140)은 화합물-단백질 복합체 (compound-protein complex) 자유 데이터를 입력받은 것 외에 상기 S100 내지 S400 단계와 동일한 절차로 제2 상호작용 행렬을 생성하여 제2 인공지능 모델을 학습시킬 수 있다.

【0100】 S600 단계는 학습한 제2 인공지능 모델(200)에서 비공유 결합 상호 작용 및 결합 친화도를 예측하는 단계일 수 있다. 일 실시예에서, 제1 상호작용 행렬과 제2 상호작용 행렬을 비교하여 손실 함수를 산출하는 단계를 포함할 수 있다.

【0101】 구체적으로, 제2 인공지능 모델(200)에서 총 2단계의 최적화 과정에서 2개의 손실함수를 산출할 수 있다. 첫번째 손실함수는 하기 수학식 8 및 수학식 9로 나타낼 수 있다.

【0102】 【수학식 8】

$$\mathcal{L}_{stage_1} = \mathcal{L}_{inter} + \mathcal{L}_{hint}$$

【0103】 【수학식 9】

$$\mathcal{L}_{inter} = \frac{1}{N} \sum_{n=1}^N \sum_{i=1}^{N_a} \sum_{j=1}^{L_p} -\|\hat{M}_{ij}^T \log M_{ij}^S + (1 - \hat{M}_{ij}^T) \log (1 - M_{ij}^S)\|_n$$

$$\mathcal{L}_{hint} = 0.5 \left(\frac{1}{N} \sum_{n=1}^N \Phi_n \|\psi_T^c(I; W_{hint}^c) - \psi_S^c(I; W_{guide}^c)\|_n^2 + \frac{1}{N} \sum_{n=1}^N \Phi_n \|\psi_T^p(I; W_{hint}^p) - \psi_S^p(I; W_{guide}^p)\|_n^2 \right)$$

【0104】

【0105】 여기서, M_{ij}^S 는 제2 인공지능 모델(200)에서 예측된 i번째 원소와 j번째 잔기 사이의 상호작용을, \hat{M}_{ij}^T 는 제1 인공지능 모델(100)에서 획득한 예측된 i

번째 원소와 j번째 잔기 사이의 상호작용을, W_{hint} 및 W_{guide} 는 각각 화합물 속성 벡터를 기초로 어텐션 값을 산출하기 위한 레이어의 파라미터, 및 단백질 속성 벡터를 기초로 어텐션 값을 산출하기 위한 레이어의 파라미터를, ψ_T 및 ψ_S 는 각각 제1 인공지능 모델(100) 및 제2 인공지능 모델(200)의 심층 신경 기능(deep neural function)을, Φ_n 는 제1 인공지능 모델(100)의 예측에 대한 신뢰도를 양자화하는 파라미터를 나타낼 수 있다.

【0106】 제2 인공지능 모델(200)에서 산출된 두번째 손실함수는 하기 수학적 식 10으로 나타낼 수 있다.

【0107】 【수학적 식 10】

$$\mathcal{L}_{stage_2} = \mathcal{L}_{reg} + \alpha_S \mathcal{L}_{imit}$$

$$\mathcal{L}_{reg} = \frac{1}{N} \sum_{n=1}^N \|P_S - P_{gt}\|_n^2$$

$$\mathcal{L}_{imit} = \frac{1}{N} \sum_{n=1}^N \Phi_n \|P_S - P_T\|_n^2$$

【0108】 여기서, α_S 는 최종 예측된 친화도를 제어하는 가중치 매개변수를, P_S , P_T , 및 P_{gt} 는 각각 제2 인공지능 모델(200), 제1 인공지능 모델(100)의 예측, 및 실제 레이블을 나타낼 수 있다.

【0110】 도 5는 일 실시예에 따른 화합물-단백질의 결합 친화도를 예측하는 모델들 간의 성능을 비교한 그래프이다. 도 5에서 Blendnet(s)은 본 개시에 따른 제2 인공지능 모델(200)을 나타낸다. 도 5를 참조하면, 새로운 단백질 기준 분할, 새로운 화합물 기준 분할, 및 무작위 분할 모델 평가 모두에서 본 발명의 예측 방법은 다른 예측 모델과 비교하여 우수한 예측 성능을 보여주었다.

【0112】 도 6은 일 실시예에 따른 컴퓨팅 장치의 하드웨어 구성을 나타내는 블록도이다.

【0113】 컴퓨팅 장치(1000)는 메모리(1100), 프로세서(1200), 통신 모듈(1300) 및 입출력 인터페이스(1400)를 포함할 수 있으며, 도 6에 도시된 바와 같이, 컴퓨팅 장치(1000)는 통신 모듈(1300)을 이용하여 네트워크를 통해 정보 및/또는 데이터를 통신할 수 있도록 구성될 수 있다.

【0114】 메모리(1100)는 비-일시적인 임의의 컴퓨터 판독 가능한 기록매체를 포함할 수 있다. 일 실시예에 따르면, 메모리(1100)는 RAM(random access memory), ROM(read only memory), 디스크 드라이브, SSD(solid state drive), 플래시 메모리(flash memory) 등과 같은 비소멸성 대용량 저장 장치(permanent mass storage device)를 포함할 수 있다. 다른 예로서, ROM, SSD, 플래시 메모리, 디스크 드라이브 등과 같은 비소멸성 대용량 저장 장치는 메모리와는 구분되는 별도의 영구 저장

장치로서 컴퓨팅 장치(1000)에 포함될 수 있다. 또한, 메모리(1100)에는 제1 인공지능 모델(100) 및 제2 인공지능 모델(200)이 저장될 수 있으며, 운영체제와 적어도 하나의 프로그램 코드가 저장될 수 있다.

【0115】 이러한 소프트웨어 구성요소들은 메모리(1100)와는 별도의 컴퓨터에서 판독 가능한 기록매체로부터 로딩될 수 있다. 이러한 별도의 컴퓨터에서 판독 가능한 기록매체는 이러한 컴퓨팅 장치(1000)에 직접 연결가능한 기록 매체를 포함할 수 있는데, 예를 들어, 플로피 드라이브, 디스크, 테이프, DVD/CD-ROM 드라이브, 메모리 카드 등의 컴퓨터에서 판독 가능한 기록매체를 포함할 수 있다. 다른 예로서, 소프트웨어 구성요소들은 컴퓨터에서 판독 가능한 기록매체가 아닌 통신 모듈(1300)을 통해 메모리(1100)에 로딩될 수도 있다. 예를 들어, 적어도 하나의 프로그램은 개발자들 또는 애플리케이션의 설치 파일을 배포하는 파일 배포 시스템이 통신 모듈(1300)을 통해 제공하는 파일들에 의해 설치되는 컴퓨터 프로그램에 기반하여 메모리(1100)에 로딩될 수 있다.

【0116】 프로세서(1200)는 기본적인 산술, 로직 및 입출력 연산을 수행함으로써, 컴퓨터 프로그램의 명령을 처리하도록 구성될 수 있다. 명령은 메모리(1100) 또는 통신 모듈(1300)에 의해 다른 사용자 단말(미도시) 또는 다른 외부 시스템으로 제공될 수 있다.

【0117】 통신 모듈(1300)은 네트워크를 통해 사용자 단말(미도시)과 컴퓨팅 장치(1000)가 서로 통신하기 위한 구성 또는 기능을 제공할 수 있으며, 컴퓨팅 장치(1000)가 외부 시스템(일례로 별도의 클라우드 시스템 등)과 통신하기 위한 구성

또는 기능을 제공할 수 있다. 일례로, 컴퓨팅 장치(1000)의 프로세서(1200)의 제어에 따라 제공되는 제어 신호, 명령, 데이터 등이 통신 모듈(1300)과 네트워크를 거쳐 사용자 단말 및/또는 외부 시스템의 통신 모듈을 통해 사용자 단말 및/또는 외부 시스템으로 전송될 수 있다.

【0118】 또한, 컴퓨팅 장치(1000)의 입출력 인터페이스(1400)는 컴퓨팅 장치(1000)와 연결되거나 컴퓨팅 장치(1000)가 포함할 수 있는 입력 또는 출력을 위한 장치(미도시)와의 인터페이스를 위한 수단일 수 있다. 도 6에서는 입출력 인터페이스(1400)가 프로세서(1200)와 별도로 구성된 요소로서 도시되었으나, 이에 한정되지 않으며, 입출력 인터페이스(1400)가 프로세서(1200)에 포함되도록 구성될 수 있다. 컴퓨팅 장치(1000)는 도 6의 구성요소들보다 더 많은 구성요소들을 포함할 수 있다. 그러나, 대부분의 종래기술적 구성요소들을 명확하게 도시할 필요성은 없다.

【0119】 컴퓨팅 장치(1000)의 프로세서(1200)는 복수의 사용자 단말 및/또는 복수의 외부 시스템으로부터 수신된 정보 및/또는 데이터를 관리, 처리 및/또는 저장하도록 구성될 수 있다.

【0120】 상술된 방법 및/또는 다양한 실시예들은, 디지털 전자 회로, 컴퓨터 하드웨어, 펌웨어, 소프트웨어 및/또는 이들의 조합으로 실현될 수 있다. 본 개시의 다양한 실시예들은 데이터 처리 장치, 예를 들어, 프로그래밍 가능한 하나 이상의 프로세서 및/또는 하나 이상의 컴퓨팅 장치에 의해 실행되거나, 컴퓨터 판독 가능한 기록 매체 및/또는 컴퓨터 판독 가능한 기록 매체에 저장된 컴퓨터 프로그램으로 구현될 수 있다. 상술된 컴퓨터 프로그램은 컴파일된 언어 또는 해석된 언어

를 포함하여 임의의 형태의 프로그래밍 언어로 작성될 수 있으며, 독립 실행형 프로그램, 모듈, 서브 루틴 등의 임의의 형태로 배포될 수 있다. 컴퓨터 프로그램은 하나의 컴퓨팅 장치, 동일한 네트워크를 통해 연결된 복수의 컴퓨팅 장치 및/또는 복수의 상이한 네트워크를 통해 연결되도록 분산된 복수의 컴퓨팅 장치를 통해 배포될 수 있다.

【0121】 한편, 본 명세서에 개시된 실시 예들은 컴퓨터에 의해 실행 가능한 명령어를 저장하는 기록매체의 형태로 구현될 수 있다. 명령어는 프로그램 코드의 형태로 저장될 수 있으며, 프로세서에 의해 실행되었을 때, 프로그램 모듈을 생성하여 개시된 실시 예들의 동작을 수행할 수 있다. 기록매체는 컴퓨터로 읽을 수 있는 기록매체로 구현될 수 있다. 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터에 의하여 해독될 수 있는 명령어가 저장된 모든 종류의 기록 매체를 포함할 수 있다. 예를 들어, ROM, RAM, 자기테이프, 자기디스크, 플래시 메모리, 광 데이터 저장 장치 등이 있을 수 있다.

【0122】 상술된 내용은 본 개시를 실시하기 위한 구체적인 실시 예들이다. 본 개시는 상술된 실시 예들뿐만 아니라, 단순히 설계 변경되거나 용이하게 변경할 수 있는 실시 예들 또한 포함할 것이다. 또한, 본 개시는 상술된 실시 예들을 이용하여 용이하게 변형하여 실시할 수 있는 기술들도 포함할 것이다. 따라서, 본 개시의 범위는 상술된 실시 예들에 국한되어 정해져서는 안 되며 후술하는 특허청구범위뿐만 아니라 본 개시의 특허청구범위와 균등한 것들에 의해 정해져야 할 것이다.

【청구범위】**【청구항 1】**

적어도 하나 이상의 프로세서에 의해 수행되는 화합물-단백질 결합 친화도 예측 방법에 있어서,

상호작용하는 화합물 데이터 및 단백질 데이터를 입력받는 단계;

상기 입력된 화합물 데이터 및 상기 단백질의 데이터를 기초로 화합물 속성 벡터 및 단백질 속성 벡터를 생성하는 단계;

상기 화합물의 속성 벡터와 상기 단백질의 속성 벡터를 기초로 어텐션 값을 산출하는 단계;

상기 어텐션 값을 기초로 제1 상호작용 행렬을 생성하는 단계;

상기 제1 상호작용 행렬을 학습 데이터로 하여 화합물-단백질의 비공유 상호작용(non-covalent interaction) 및 결합 친화도(binding affinity)을 예측하도록 제1 인공지능 모델을 학습시키는 단계; 및

상기 제1 인공지능 모델의 출력값을 기초로 화합물-단백질의 비공유 상호작용 및 결합 친화도를 예측하는 단계;를 포함하는

화합물-단백질 결합 친화도 예측 방법.

【청구항 2】

제1 항에 있어서,

상기 화합물 데이터를 입력받는 단계는

화합물 구조 그래프를 생성하는 단계를 포함하며,

상기 화합물 구조 그래프는 원자의 속성을 노드로, 결합(bond)의 속성을 엣지로 한 것인,

화합물-단백질 결합 친화도 예측 방법.

【청구항 3】

제1 항에 있어서,

상기 화합물의 속성 벡터와 상기 단백질의 속성 벡터를 기초로 어텐션 값을 산출하는 단계는,

상기 화합물의 속성 벡터와 상기 단백질의 속성 벡터를 기초로 각각의 키(key), 벨류(value), 및 쿼리(query)를 산출하는 단계; 및

상기 산출된 화합물의 쿼리와 단백질의 쿼리를 교차하거나, 화합물의 키 및 벨류와 단백질의 키 및 벨류를 교차해 서브 어텐션 레이어에 제공하는 단계; 를 포함하는

화합물-단백질 결합 친화도 예측 방법.

【청구항 4】

제3 항에 있어서,

상기 화합물의 속성 벡터와 상기 단백질의 속성 벡터를 기초로 어텐션 값을 산출하는 단계는,

상기 서브 어텐션 레이어의 출력값을 셀프 어텐션 레이어에 제공하는 단계;
및

상기 셀프 어텐션 레이어의 출력값을 피드-포워드 레이어(feed-forward layer)에 제공하는 단계; 를 더 포함하는,

화합물-단백질 결합 친화도 예측 방법.

【청구항 5】

제1 항에 있어서,

상기 제1 인공지능 모델을 학습시키는 단계는,

상기 화합물의 하나 이상의 원소 및 상기 단백질의 하나 이상의 잔기 사이의 상호작용 점수를 산출하는 단계;

상기 점수에서 잠재변수를 추출하는 단계; 및

상기 잠재변수를 완전 연결 레이어(fully connected layer)에 제공하는 단계;를 포함하는

화합물-단백질 결합 친화도 예측 방법.

【청구항 6】

제1 항에 있어서,

상기 제1 인공지능 모델의 출력값은

제1 인공지능 모델에서 예측된 결합 친화도, 상기 화합물의 속성 벡터, 상기 단백질의 속성 벡터, 및 상기 제1 상호작용 행렬을 포함하는

화합물-단백질 결합 친화도 예측 방법.

【청구항 7】

제1 항에 있어서,

상기 제1 인공지능 모델의 출력값을 기초로 화합물-단백질의 비공유 상호작용(non-covalent interaction) 및 결합 친화도를 예측하는 단계는,

상기 제1 인공지능 모델의 출력값을 제2 인공지능 모델에 입력해 지식 증류를 수행하는 단계; 및

상기 제2 인공지능 모델에서 비공유 상호작용(non-covalent interaction) 및 결합 친화도를 예측을 수행하는 단계를 포함하고,

상기 제1 인공지능 모델이 입력받은 데이터는 화합물-단백질 복합체(compound-protein complex)에 기초한 데이터인 것인,

화합물-단백질 결합 친화도 예측 방법.

【청구항 8】

제 7항에 있어서,

상기 지식 증류를 수행하는 단계는,

제1 상호작용 행렬과 상기 제2 인공지능 모델에서 생성한 제2 상호작용 행렬을 비교하여 손실 함수를 산출하는 단계를 포함하는,

화합물-단백질 결합 친화도 예측 방법.

【청구항 9】

제1 항 내지 제 8항 중 어느 한 항에 기재된 화합물-단백질 결합 친화도 예측 방법을 실행시키도록 판독 가능한 기록 매체에 기록된 컴퓨터 프로그램.

【청구항 10】

컴퓨팅 장치로서,

통신 모듈;

메모리; 및

상기 메모리와 연결되고, 상기 메모리에 포함된 컴퓨터 판독 가능한 적어도 하나의 프로그램을 실행하도록 구성된 적어도 하나의 프로세서;

를 포함하고,

상기 적어도 하나의 프로그램은,

상호작용하는 화합물 및 단백질의 구조를 입력받고,

상기 입력된 화합물 구조 및 상기 단백질의 구조를 기초로 화합물 속성 벡터
및 단백질 속성 벡터를 생성하고,

상기 화합물의 속성 벡터와 상기 단백질의 속성 벡터를 기초로 어텐션 값을
산출하고,

상기 어텐션 값을 기초로 제1 상호작용 행렬을 생성하고,

상기 제1 상호작용 행렬을 학습 데이터로 하여 화합물-단백질의 비공유 상호
작용(non-covalent interaction) 및 결합 친화도(binding affinity)을 예측하도록
제1 인공지능 모델을 학습시키고,

상기 제1 인공지능 모델의 출력값을 기초로 화합물-단백질의 비공유 상호작
용 및 결합 친화도를 예측하는 컴퓨팅 장치.

【요약서】

【요약】

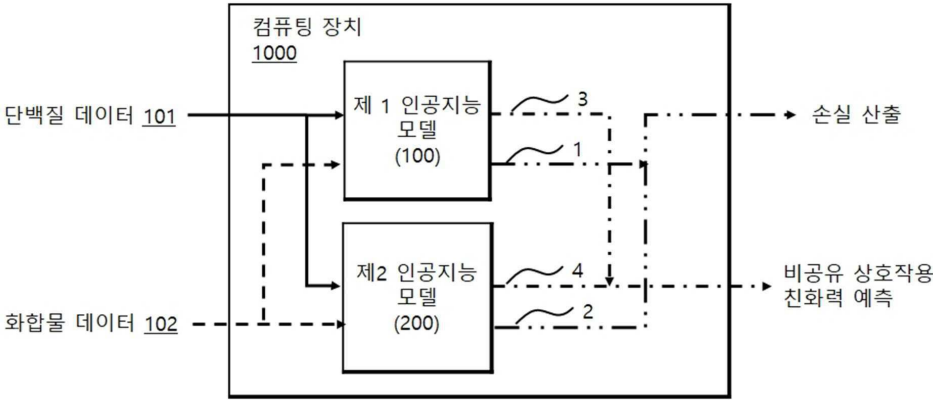
적어도 하나 이상의 프로세서에 의해 수행되는 화합물-단백질 결합 친화도 예측 방법에 있어서, 상호작용하는 화합물 데이터 및 단백질 데이터를 입력받는 단계; 상기 입력된 화합물 데이터 및 상기 단백질의 데이터를 기초로 화합물 속성 벡터 및 단백질 속성 벡터를 생성하는 단계; 상기 화합물의 속성 벡터와 상기 단백질의 속성 벡터를 기초로 어텐션 값을 산출하는 단계; 상기 어텐션 값을 기초로 제1 상호작용 행렬을 생성하는 단계; 상기 제1 상호작용 행렬을 학습 데이터로 하여 화합물-단백질의 비공유 상호작용(non-covalent interaction) 및 결합 친화도(binding affinity)을 예측하도록 제1 인공지능 모델을 학습시키는 단계; 및 상기 제1 인공지능 모델의 출력값을 기초로 화합물-단백질의 비공유 상호작용 및 결합 친화도를 예측하는 단계;를 포함하는 화합물-단백질 결합 친화도 예측 방법을 제공할 수 있다.

【대표도】

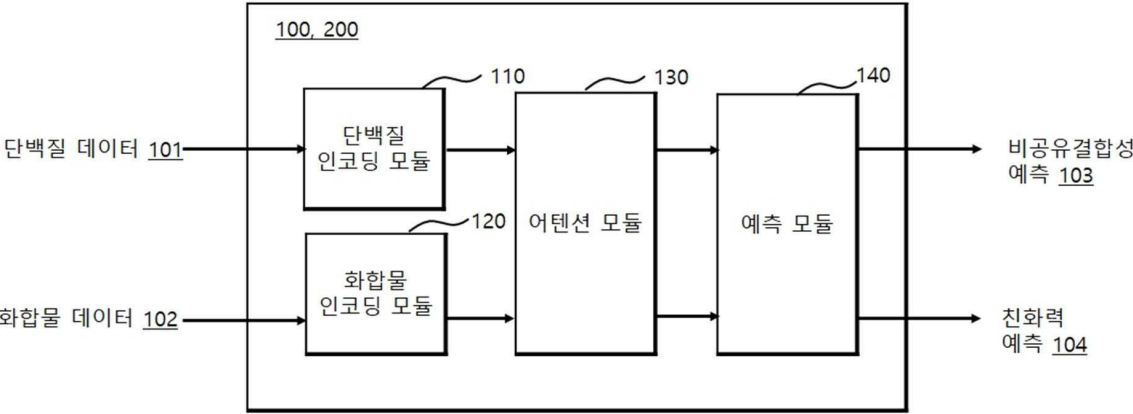
도 1

【도면】

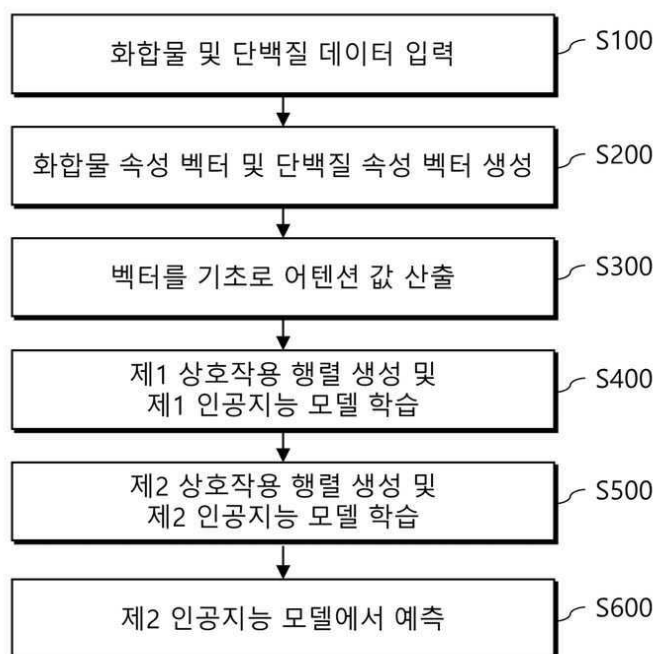
【도 1】



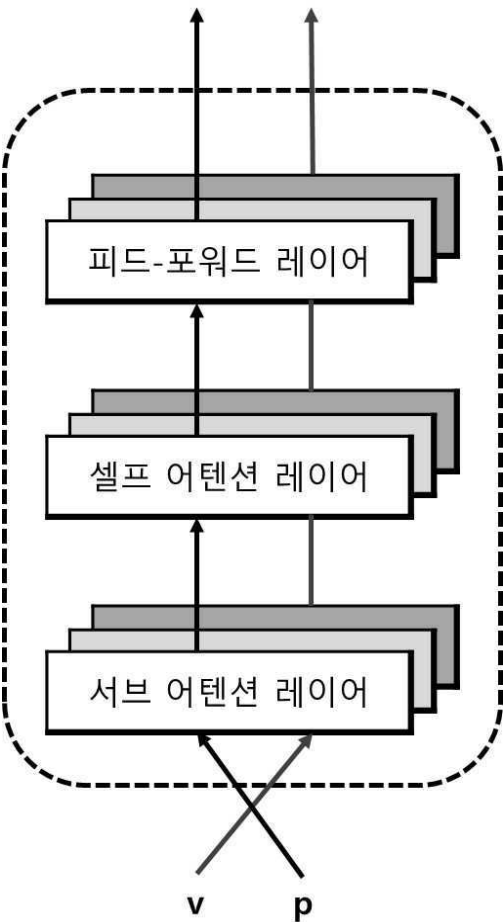
【도 2】



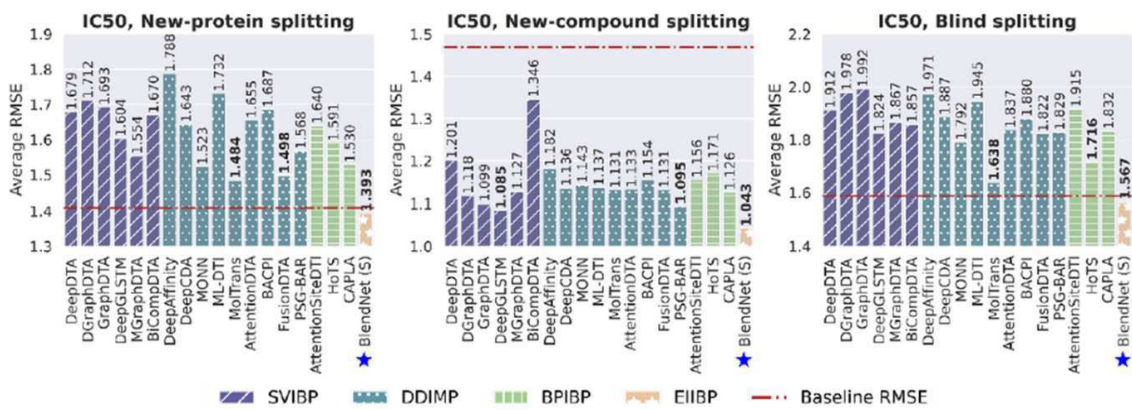
【도 3】



【도 4】



【도 5】



【도 6】

