【서지사항】

【서류명】 특허출원서

【참조번호】 SDP20244687

【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】 단백질 서열 기반의 리간드 결합 부위 예측 방법 및 장치

【발명의 영문명칭】 METHOD AND APPARATUS FOR PREDICTING LIGAND BINDING SITE BASED ON PROTEIN SEQUENCE

【발명자】

【성명】 박상현

【성명의 영문표기】 SANGHYUN PARK

【주민등록번호】 670101-1XXXXXX

【우편번호】 08004

【주소】 서울특별시 양천구 오목로 300, 204동 3701호

【발명자】

【성명】 서상민

【성명의 영문표기】 SANGMIN SEO

【주민등록번호】 930507-1XXXXXX

【우편번호】 63272

【주소】 제주특별자치도 제주시 고마로 44, 801호

【발명자】

【성명】 최종환

【성명의 영문표기】 JONGHWAN CHOI

【주민등록번호】 910226-1XXXXXX

【우편번호】 24250

【주소】 강원특별자치도 춘천시 후석로 459, 104동 1503호

【발명자】

【성명】 최승연

【성명의 영문표기】 SEUNGYEON CHOI

【주민등록번호】 970828-1XXXXXX

【우편번호】 07213

【주소】 서울특별시 영등포구 당산로45길 7-3, 101동 808호

【발명자】

【성명】 이지은

【성명의 영문표기】 JIEUN LEE

【주민등록번호】 960912-2XXXXXX

【우편번호】 03716

【주소】 서울특별시 서대문구 동교로 291, 101동 1004호

【출원언어】 국어

【심사청구】 청구

【공지예외적용대상증명서류의 내용】

【공개형태】 논문

【공개일자】 2023.10.18

【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 적응증 최적화 플랫폼 개발 및 혁신신약 발굴

【과제수행기관명】 연세대학교

【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】 47 면 0 원

【우선권주장료】	0	건	0 원
【심사청구료】	13	항	829,000 원
【합계】	875,000원		
【감면사유】	전담조직(50%감면)[1]		
【감면후 수수료】	437,500 원		
【첨부서류】	1.공지예외적용대상(신규성상실의예외, 출원시의특례)규정을 적용받기 위한 증명서류[SDP2024-4687_공지예외주장]_1통		

1 : 공지예외적용대상(신규성상실의예외,_출원시의특례)규정을_적용받기_위한_증명서류
PDF 파일 첨부

【발명의 설명】

【발명의 명칭】

단백질 서열 기반의 리간드 결합 부위 예측 방법 및 장치{METHOD AND APPARATUS FOR PREDICTING LIGAND BINDING SITE BASED ON PROTEIN SEQUENCE}

【기술분야】

【0001】 본 개시는 단백질 서열 기반의 리간드 결합 부위 예측 방법 및 장치에 관한 것으로, 구체적으로, 리간드가 결합되는 아미노산 잔기를 예측하기 위한 방법 및 장치에 관한 것이다.

【발명의 배경이 되는 기술】

【0002】 단백질 및 리간드(ligand)의 상호작용은 많은 생물학적 과정에서 중요한 역할을 하며, 성공적인 약물 설계 등을 위해 단백질의 리간드 결합 부위를 정밀하게 예측하는 것이 요구된다. 리간드 결합 부위를 예측하기 위한 방법으로, 단백질의 시각적인 3차원 구조를 이용하는 3D 구조 기반의 기술들이 사용되고 있으나, 단백질의 3차원 구조를 결정하는 것은 높은 자원이 소모되는 문제가 있다.

【발명의 내용】

【해결하고자 하는 과제】

【0003】 본 개시는 상기와 같은 문제점을 해결하기 위한 단백질 서열 기반의 리간드 결합 부위 예측 방법, 컴퓨터 판독 가능 매체에 저장된 컴퓨터 프로그램, 컴퓨터 프로그램이 저장된 컴퓨터 판독 가능 매체 및 장치(시스템)를 제공한다.

【과제의 해결 수단】

【0004】 본 개시는 방법, 장치(시스템), 컴퓨터 판독 가능 매체에 저장된 컴퓨터 프로그램 또는 컴퓨터 프로그램이 저장된 컴퓨터 판독 가능 매체를 포함한 다양한 방식으로 구현될 수 있다.

【0005】 본 개시의 일 실시예에 따르면, 적어도 하나의 프로세서에 의해 수행되는 단백질 서열 기반의 리간드 결합 부위 예측 방법은, 단백질 서열을 입력받는 단계, 입력된 단백질 서열을 기초로 아미노산 레벨 임베딩 벡터 및 단백질 레벨 임베딩 벡터를 생성하는 단계 및 생성된 아미노산 레벨 임베딩 벡터 및 단백질 레벨 임베딩 벡터를 이용하여 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정하는 단계를 포함한다.

【0006】 본 개시의 일 실시예에 따르면, 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정하는 단계는, 아미노산 레벨 임베딩 벡터를 이용하여 잔기 임베딩 벡터를 생성하는 단계, 생성된 잔기 임베딩 벡터를 학습된 인공지능 모델에 제공하여 인접한 아미노산 잔기 사이의 지역적 특징을 나타내는 키 및 밸류를 산출하는 단계, 단백질 레벨 임베딩 벡터를 이용하여 쿼리를 산출하는 단계, 산출된 키, 밸류 및 쿼리를 어텐션 모듈에 제공하여 어텐션 값을 산출하는 단계 및 산출된 어텐션 값을 기초로 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정하는 단계를 포함한다.

【0007】본 개시의 일 실시예에 따르면, 아미노산 레벨 임베딩 벡터를 이용하여 잔기 임베딩 벡터를 생성하는 단계는, 단백질 서열을 구성하는 단백질 체인과 연관된 체인 임베딩 벡터를 생성하는 단계, 단백질 서열을 구성하는 각각의 아미노산의 위치와 연관된 위치 임베딩 벡터를 생성하는 단계 및 생성된 체인 임베딩 벡터 및 위치 임베딩 벡터를 아미노산 레벨 임베딩 벡터에 더하여 잔기 임베딩 벡터를 생성하는 단계를 포함한다.

【0008】본 개시의 일 실시예에 따르면, 단백질 레벨 임베딩 벡터를 이용하여 쿼리를 산출하는 단계는, 단백질 서열을 구성하는 단백질 체인과 연관된 체인 임베딩 벡터를 생성하는 단계, 단백질 서열을 구성하는 각각의 아미노산의 위치와 연관된 위치 임베딩 벡터를 생성하는 단계 및 생성된 체인 임베딩 벡터 및 위치 임베딩 벡터를 단백질 레벨 임베딩 벡터에 더하여 쿼리를 산출하는 단계를 포함한다.

【0009】본 개시의 일 실시예에 따르면, 산출된 어텐션 값을 기초로 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정하는 단계는, 어텐션 값 및 단백질 레벨 임베딩 벡터를 연결하여 단백질 서열에 대응하는 대표 벡터를 생성하는 단계 및 생성된 대표 벡터를 결합 부위 여부를 판정하도록 학습된 분류기에 제공하여 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정하는 단계를 포함한다.

【0010】본 개시의 일 실시예에 따르면, 분류기는, 4개의 은닉층을 갖는 완전 연결 레이어를 포함한다.

【0011】본 개시의 일 실시예에 따르면, 분류기는, 가중 교차 엔트로피 손실

함수를 기초로 학습된다.

【0012】 본 개시의 일 실시예에 따르면, 인공지능 모델은, 합성곱 신경망 기반의 모델을 포함한다.

【0013】 본 개시의 일 실시예에 따르면, 합성곱 신경망은, 복수의 합성곱 신경망 블록을 포함하는 1차원 합성곱 신경망이다.

【0014】 본 개시의 일 실시예에 따르면, 복수의 합성곱 신경망 블록에 포함된 각각의 합성곱 신경망 블록은 서로 다른 커널 폭을 갖는 3개의 합성곱 신경망 레이어를 포함한다.

【0015】 본 개시의 일 실시예에 따르면, 입력된 단백질 서열을 기초로 아미노산 레벨 임베딩 벡터 및 단백질 레벨 임베딩 벡터를 생성하는 단계는, 입력된 단백질 서열을 구성하는 각각의 아미노산에 대한 임베딩을 수행하여 아미노산 레벨 임베딩 벡터를 생성하는 단계 및 생성된 아미노산 레벨 임베딩 벡터의 평균값을 기초로 단백질 레벨 임베딩 벡터를 생성하는 단계를 포함한다.

【0016】 본 개시의 일 실시예에 따른 상술된 방법을 컴퓨터에서 실행하기 위해 컴퓨터 판독 가능한 기록 매체에 저장된 컴퓨터 프로그램이 제공된다.

【0017】 본 개시의 일 실시예에 따른 컴퓨팅 장치는, 통신 모듈, 메모리 및 메모리와 연결되고, 메모리에 포함된 컴퓨터 판독 가능한 적어도 하나의 프로그램을 실행하도록 구성된 적어도 하나의 프로세서를 포함한다. 적어도 하나의 프로그램은, 단백질 서열을 입력받고, 입력된 단백질 서열을 기초로 아미노산 레벨 임베

딩 벡터 및 단백질 레벨 임베딩 벡터를 생성하고, 생성된 아미노산 레벨 임베딩 벡터 및 단백질 레벨 임베딩 벡터를 이용하여 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정하기 위한 명령어들을 포함한다.

【발명의 효과】

【0018】 본 개시의 다양한 실시예에서 컴퓨팅 장치는 단백질 서열을 입력받는 것만으로도, 인접한 아미노산 잔기 사이의 지역적인 특징뿐만 아니라, 거리 의존성을 갖는 아미노산 잔기 사이의 전역적인 특징을 모두 고려하여 결합 부위 예측 성능을 효과적으로 향상시킬 수 있다.

【0019】 본 개시의 다양한 실시예에서 아미노산 레벨 임베딩 벡터(330)와 함께 단백질 레벨 임베딩 벡터(340)를 생성하여 인공지능 모델 등의 입력으로 사용함으로써 단백질 서열 상의 지역적인 정보뿐만 아니라 전역적인 정보까지 효과적으로 활용될 수 있다.

【도면의 간단한 설명】

【0020】 본 개시의 실시예들은, 이하 설명하는 첨부 도면들을 참조하여 설명될 것이며, 여기서 유사한 참조 번호는 유사한 요소들을 나타내지만, 이에 한정되지는 않는다.

도 1은 본 개시의 일 실시예에 따른 컴퓨팅 장치의 내부 구성을 나타내는 기능적인 블록도이다.

도 2는 본 개시의 일 실시예에 따른 단백질 서열을 시각화한 예시적인 이미

지를 나타내는 도면이다.

도 3은 본 개시의 일 실시예에 따른 아미노산 레벨 임베딩 벡터 및 단백질 레벨 임베딩 벡터가 생성되는 예시를 나타내는 도면이다.

도 4는 본 개시의 일 실시예에 따른 분류기에 의해 리간드 결합 부위 예측이 수행되는 과정의 예시를 나타내는 도면이다.

도 5는 본 개시의 일 실시예에 따른 인공지능 모델의 예시적인 구조를 나타내는 도면이다.

도 6은 본 개시의 일 실시예에 따른 리간드 결합 부위를 예측하는 모델들 간의 성능을 비교한 그래프이다.

도 7은 본 개시의 일 실시예에 따른 리간드 결합 부위를 예측하는 모델들 간의 예측 결과를 시각화한 이미지이다.

도 8은 본 개시의 일 실시예에 따른 단백질 서열 기반의 리간드 결합 부위 예측 방법의 예시를 나타내는 흐름도이다.

도 9는 본 개시의 일 실시예에 따른 컴퓨팅 장치의 하드웨어 구성을 나타내는 블록도이다.

【발명을 실시하기 위한 구체적인 내용】

【0021】 이하, 본 개시의 실시를 위한 구체적인 내용을 첨부된 도면을 참조하여 상세히 설명한다. 다만, 이하의 설명에서는 본 개시의 요지를 불필요하게 흐릴 우려가 있는 경우, 널리 알려진 기능이나 구성에 관한 구체적 설명은 생략하기

로 한다.

【0022】첨부된 도면에서, 동일하거나 대응하는 구성요소에는 동일한 참조부호가 부여되어 있다. 또한, 이하의 실시예들의 설명에 있어서, 동일하거나 대응되는 구성요소를 중복하여 기술하는 것이 생략될 수 있다. 그러나, 구성요소에 관한 기술이 생략되어도, 그러한 구성요소가 어떤 실시예에 포함되지 않는 것으로 의도되지는 않는다.

【0023】개시된 실시예의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명이 완전하도록 하고, 본 발명이 통상의 기술자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것일 뿐이다.

【0024】본 명세서에서 사용되는 용어에 대해 간략히 설명하고, 개시된 실시예에 대해 구체적으로 설명하기로 한다. 본 명세서에서 사용되는 용어는 본 발명에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 관련 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 발명의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서, 본 발명에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 개시의 전반에 걸친 내용을 토대로 정의되어야 한다.

【0025】본 명세서에서의 단수의 표현은 문맥상 명백하게 단수인 것으로 특정하지 않는 한, 복수의 표현을 포함한다. 또한, 복수의 표현은 문맥상 명백하게 복수인 것으로 특정하지 않는 한, 단수의 표현을 포함한다. 명세서 전체에서 어떤 부분이 어떤 구성요소를 포함한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있음을 의미한다.

【0026】본 개시에서, "포함하다", "포함하는" 등의 용어는 특징들, 단계들, 동작들, 요소들 및/또는 구성 요소들이 존재하는 것을 나타낼 수 있으나, 이러한 용어가 하나 이상의 다른 기능들, 단계들, 동작들, 요소들, 구성 요소들 및/또는 이들의 조합이 추가되는 것을 배제하지는 않는다.

【0027】본 개시에서, 특정 구성 요소가 임의의 다른 구성 요소에 "결합", "조합", "연결" 되거나, "반응" 하는 것으로 언급된 경우, 특정 구성 요소는 다른 구성 요소에 직접 결합, 조합 및/또는 연결되거나, 반응할 수 있으나, 이에 한정되지 않는다. 예를 들어, 특정 구성 요소와 다른 구성 요소 사이에 하나 이상의 중간 구성 요소가 존재할 수 있다. 또한, 본 발명에서 "및/또는"은 열거된 하나 이상의 항목의 각각 또는 하나 이상의 항목의 적어도 일부의 조합을 포함할 수 있다.

【0028】본 개시에서, "제1", "제2" 등의 용어는 특정 구성 요소를 다른 구성 요소와 구별하기 위해 사용되는 것으로, 이러한 용어에 의해 상술된 구성 요소가 제한되진 않는다. 예를 들어, "제1" 구성 요소는 "제2" 구성 요소와 동일하거나 유사한 형태의 요소일 수 있다.

【0029】본 개시에서 "쪈부", "쪈모듈" 등의 용어는 적어도 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 개시에서, 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 개시에서, 기 정의된 동작 규칙 또는 인공지능 모델은 기계학습을 수행하도록 구성될 수 있다. 여기서, 기계학습을 수행하도록 구성된다는 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 인공신경망)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 개시에서, "단백질 서열(protein sequence)"은 단백질을 아미노산의 배열한 것을 지칭할 수 있다. 본 개시에서, 단백질, 단백질 서열, 아미노산 등은 텍스트 데이터(text data)로 표현될 수 있다.

【0035】 본 개시에서, "리간드 결합 부위(ligand binding site)"는 단백질과 반응하기 위해 리간드(ligand)가 결합할 수 있는 위치 및/또는 아미노산을 지칭할 수 있다.

【0036】 본 개시에서, "임베딩 모듈(embedding module)"은 단백질 서열 등에 관한 텍스트 데이터를 컴퓨터가 이행할 수 있는 숫자 형태인 벡터로 변환하는 임의의 구성을 지칭할 수 있다.

【0037】 본 개시에서, "어텐션 모듈(attention module)"은 어텐션 함수를 이용하여 어텐션 연산을 수행하는 임의의 구성을 지칭할 수 있다. 예를 들어, 어텐션

함수는 쿼리(query)에 대해 모든 키(key)와 유사도를 구하고, 이 유사도를 키와 매핑되어 있는 각각의 밸류(value)에 반영한 후, 유사도가 반영된 밸류를 모두 리턴하여 어텐션 값(attention value)을 반환할 수 있다.

【0038】 도 1은 본 개시의 일 실시예에 따른 컴퓨팅 장치(100)의 내부 구성을 나타내는 기능적인 블록도이다. 일 실시예에 따르면, 컴퓨팅 장치(100)는 단백질 서열(102)을 입력받아 해당 단백질 서열(102)에 포함된 리간드 결합 부위(104)를 예측하기 위한 임의의 장치를 지칭할 수 있다. 예를 들어, 컴퓨팅 장치(100)는 임베딩 모듈(110), 인공지능 모델(120), 어텐션 모듈(130) 및 분류기(140) 등을 포함할 수 있으나, 이에 한정되지 않는다.

【0039】 일 실시예에 따르면, 컴퓨팅 장치(100)는 리간드 결합 부위(104)를 예측하기 위한 단백질 서열(102)을 입력받거나 수신할 수 있다. 여기서, 단백질 서열(102)은 20 종류의 아미노산의 배열로 구성될 수 있으며, 복수의 단백질 체인(protein chain)을 포함할 수 있다. 예를 들어, 단백질 서열(102)은 20 종류의 아미노산의 배열에 따라 상이한 속성을 가질 수 있다.

【0040】 단백질 서열(102)을 입력받는 경우, 임베딩 모듈(110)은 해당 단백질 서열(102)을 인공지능 모델(120) 등의 입력으로 사용하기 위해 임베딩(embedding)을 수행할 수 있다. 여기서, 임베딩은 데이터를 수치화하기 위해 자연어를 특정 수치 값을 갖는 벡터로 변환하는 것을 나타낼 수 있다. 즉, 임베딩 모듈(110)은 단백질 서열(102)에 대한 임베딩을 수행하여 특정 수치 값을 갖는 벡터들을 생성할 수 있다.

【0041】일 실시예에 따르면, 임베딩 모듈(110)은 단백질 서열(102)을 기초로 아미노산 레벨(amino acid-level)의 벡터들과 단백질 레벨(protein-level)의 벡터들을 생성할 수 있다. 예를 들어, 임베딩 모듈(110)은 단백질 서열(102)을 구성하는 각각의 아미노산에 대한 임베딩을 수행하여 아미노산 레벨 임베딩 벡터를 생성할 수 있다. 또한, 임베딩 모듈(110)은 생성된 아미노산 레벨 임베딩 벡터의 평균값을 기초로 단백질 레벨 임베딩 벡터를 생성할 수 있다.

【0042】이와 같이 생성된 아미노산 레벨 임베딩 벡터는 인공지능 모델(120)에 제공될 수 있다. 이 때, 아미노산 레벨 임베딩 벡터는 잔기 임베딩 벡터(residue embedding vector)로 변환되어 인공지능 모델(120)에 제공될 수 있다. 예를 들어, 임베딩 모듈(110)은 단백질 서열(102)을 구성하는 단백질 체인과 연관된 체인 임베딩 벡터(chain embedding vector)를 생성하고, 단백질 서열(102)을 구성하는 각각의 아미노산의 위치와 연관된 위치 임베딩 벡터(position embedding vector)를 생성할 수 있다. 이 경우, 생성된 체인 임베딩 벡터 및 위치 임베딩 벡터가 아미노산 레벨 임베딩 벡터와 더해지는 것에 의해 잔기 임베딩 벡터가 생성될 수 있다.

【0043】일 실시예에 따르면, 인공지능 모델(120)은 잔기 임베딩 벡터를 이용하여 단백질 서열(102)에 포함된 인접한 아미노산 잔기 사이의 지역적 특징을 나타내는 키(key) 및 밸류(value)를 산출할 수 있다. 예를 들어, 인공지능 모델(120)은 합성곱 신경망(convolutional neural network; CNN) 기반의 모델을 포함할 수 있으며, 커널 폭(kernel width)을 조절하는 것에 의해 특정 범위 내의 아미노산 잔

기 사이의 지역적 특징을 포함하는 출력값을 생성할 수 있다. 이 경우, 인공지능 모델(120)의 출력값을 기초로 어텐션 모듈(130)의 입력값으로 사용되는 키 및 밸류가 산출될 수 있다.

【0044】 일 실시예에 따르면, 임베딩 모듈(110)은 체인 임베딩 벡터 및 위치 임베딩 벡터를 단백질 레벨 임베딩 벡터에 더하여 쿼리(query)를 산출할 수 있다. 이 경우, 산출된 쿼리는 키 및 밸류와 함께 어텐션 모듈(130)에 제공될 수 있다. 어텐션 모듈(130)은 키, 밸류 및 쿼리를 이용하여 단백질 서열(102)을 구성하는 각각의 아미노산과 각각의 단백질 체인 사이의 연관성을 나타내는 어텐션 값을 산출할 수 있다. 이와 같이 산출된 어텐션 값은 아미노산 잔기 사이의 긴 거리 의존성(long distance dependency)을 포착하여 산출된 값일 수 있다.

【0045】 일 실시예에 따르면, 분류기(140)는 산출된 어텐션 값을 기초로 단백질 서열(102)을 구성하는 각각의 아미노산이 결합 부위(104)인지 여부를 판정할 수 있다. 예를 들어, 어텐션 값 및 단백질 레벨 임베딩 벡터가 연결(concatenate)되어 대표 벡터가 생성될 수 있으며, 분류기(140)는 대표 벡터를 입력받아 단백질 서열(102)을 구성하는 각각의 아미노산이 결합 부위(104)인지 여부를 판정할 수 있다.

【0046】 도 1에서는 컴퓨팅 장치(100)에 포함된 각각의 기능적인 구성이 구분되어 상술되었으나, 이는 발명의 이해를 돕기 위한 것일 뿐이며, 하나의 연산 장치에서 둘 이상의 기능을 수행할 수도 있다. 이와 같은 구성에 의해, 컴퓨팅 장치(100)는 단백질 서열(102)을 입력받는 것만으로도, 인접한 아미노산 잔기 사이의

지역적인 특징뿐만 아니라, 거리 의존성을 갖는 아미노산 잔기 사이의 전역적인 특징을 모두 고려하여 결합 부위 예측 성능을 효과적으로 향상시킬 수 있다.

【0047】 도 2는 본 개시의 일 실시예에 따른 단백질 서열을 시각화한 예시적인 이미지(200)를 나타내는 도면이다. 상술된 것과 같이, 단백질 서열은 20 종류의 아미노산의 결합으로 생성될 수 있다. (a) 도면은 단백질 서열의 일부를 시각화한 것으로, M, E, N 등은 각각 하나의 아미노산과 대응되는 것일 수 있다. 상술된 것과 같이, 단백질 서열을 입력받는 경우, 컴퓨팅 장치(도 1의 100)는 해당 단백질 서열을 구성하는 각각의 아미노산 및/또는 아미노산 잔기가 리간드 결합 부위인지 여부를 판정할 수 있다.

【0048】 (b) 도면에 도시된 것과 같이, 컴퓨팅 장치는 단백질 서열을 구성하는 각각의 아미노산 및/또는 아미노산 잔기가 리간드 결합 부위일 확률 값을 산출할 수 있다. 그리고 나서, 컴퓨팅 장치는 산출된 확률 값이 임계치 이상이거나, 확률 값이 높은 상위 n개의 아미노산 및/또는 아미노산 잔기를 리간드 결합 부위로 예측할 수 있다. 도시된 예에서, 컴퓨팅 장치는 E, N, F, I 및 G에 대응하는 아미노산 및/또는 아미노산 잔기를 리간드 결합 부위로 예측할 수 있다.

【0049】 도 3은 본 개시의 일 실시예에 따른 아미노산 레벨 임베딩 벡터(330) 및 단백질 레벨 임베딩 벡터(340)가 생성되는 예시를 나타내는 도면이다. 상술된 것과 같이, 컴퓨팅 장치(도 1의 100)는 단백질 서열(310)을 입력받고, 입력된 단백질 서열(310)을 기초로 아미노산 레벨 임베딩 벡터(330) 및 단백질 레벨 임베딩 벡터(340)를 생성할 수 있다.

【0050】일 실시예에 따르면, 단백질 서열(310)은 복수의 단백질 체인으로 형성될 수 있다. 예를 들어, 제1 단백질 체인은 AA_1^1, 쩜, AA_2^1, 쩜, AA_3^1의 아미노산을 포함하고, 제2 단백질 체인은 AA_1^2, 쩜, AA_2^2, 쩜, AA_3^2의 아미노산을 포함하며, 제3 단백질 체인은 AA_1^3, 쩜, AA_2^3, 쩜, AA_3^3의 아미노산을 포함할 수 있다. 이 경우, 각각의 단백질 체인에 포함된 아미노산(예: 아미노산과 연관된 데이터)을 인공지능 모델 등의 입력으로 사용하기 위한 임베딩이 수행될 수 있다.

【0051】일 실시예에 따르면, 각각의 아미노산과 연관된 데이터가 언어 모델(320)에 제공되는 것에 의해 임베딩이 수행될 수 있다. 여기서, 언어 모델(320)은 BERT(bidirectional encoder representations from transformers; BERT) 기반의 모델로서, 예를 들어, 대규모의 단백질 서열 데이터로 사전 학습된 ProtTrans 모델을 포함할 수 있다. 즉, 언어 모델(320)은 아미노산과 연관된 데이터가 입력되는 경우, 해당 아미노산과 연관된 아미노산 레벨의 임베딩 벡터를 생성할 수 있다.

【0052】일 실시예에 따르면, 언어 모델(320)에 의해 생성되는 아미노산 레벨 임베딩 벡터(330)는 다음의 수학식 1과 같이 표현될 수 있다.

【0054】**【수학식 1】**

$$X_A \in R^{L \times d}$$

【0056】 여기서, X_A는 아미노산 레벨 임베딩 벡터(330)를 나타내고, L은 최대 단백질 서열 길이를 나타내고, d는 은닉 크기를 나타낼 수 있다. 도시된 예에서, AA_1^1, ・・・, AA_2^1, ・・・, AA_3^1의 아미노산에 대응하여 E_1^1, ・・・, E_2^1, ・・・, E_3^1의 임베딩 벡터가 생성되고, AA_1^2, ・・・, AA_2^2, ・・・, AA_3^2의 아미노산에 대응하여 E_1^2, ・・・, E_2^2, ・・・, E_3^2의 임베딩 벡터가 생성되고, AA_1^3, ・・・, AA_2^3, ・・・, AA_3^3의 아미노산에 대응하여 E_1^3, ・・・, E_2^3, ・・・, E_3^3의 임베딩 벡터가 생성될 수 있다. 이 때, 생성된 임베딩 벡터들이 병합되어 아미노산 레벨 임베딩 벡터(330)가 생성될 수 있다.

【0057】 추가적으로, 단백질 레벨 임베딩 벡터(340)는 다음의 수학식 2와 같이 표현될 수 있다.

【0059】 **【수학식 2】**

$$X_p \in R^d$$

【0061】 여기서, X_p는 단백질 레벨 임베딩 벡터(340)를 나타낼 수 있다. 일 실시예에 따르면, 단백질 레벨 임베딩 벡터(340)는 아미노산 레벨 임베딩 벡터(330)의 평균값을 기초로 생성될 수 있다. 이 경우, 단백질 레벨 임베딩 벡터(34

0)를 생성하기 위한 평균값은 각각의 단백질 체인 별로 수행될 수 있다. 예를 들어, E_1^1, 쪋, E_2^1, 쪋, E_3^1의 임베딩 벡터의 평균값을 기초로 제1 단백질 레벨 임베딩 벡터가 생성되고, E_1^2, 쪋, E_2^2, 쪋, E_3^2의 임베딩 벡터의 평균값을 기초로 제2 단백질 레벨 임베딩 벡터가 생성되고, E_1^3, 쪋, E_2^3, 쪋, E_3^3의 임베딩 벡터의 평균값을 기초로 제3 단백질 레벨 임베딩 벡터가 생성될 수 있다. 이와 같이 생성된 제1 내지 제3 단백질 레벨 임베딩 벡터 등이 병합되어 최종적인 단백질 레벨 임베딩 벡터(340)가 생성할 수 있다.

【0062】 도 3에서는 단백질 서열이 3개의 단백질 체인을 포함하는 것으로 도시되었으나, 이는 예시적인 것이며, 단백질 서열을 구성하는 아미노산의 수 및 단백질 체인의 수는 상이하게 결정될 수 있다. 이와 같은 구성에 의해, 아미노산 레벨 임베딩 벡터(330)와 함께 단백질 레벨 임베딩 벡터(340)를 생성하여 인공지능 모델 등의 입력으로 사용함으로써 단백질 서열 상의 지역적인 정보뿐만 아니라 전역적인 정보까지 효과적으로 활용될 수 있다.

【0063】 도 4는 본 개시의 일 실시예에 따른 분류기(460)에 의해 리간드 결합 부위 예측이 수행되는 과정의 예시를 나타내는 도면이다. 상술된 것과 같이, 컴퓨팅 장치(도 1의 100)는 아미노산 레벨 임베딩 벡터(330) 및 단백질 레벨 임베딩 벡터(340)를 이용하여 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정할 수 있다.

【0064】일 실시예에 따르면, 아미노산 레벨 임베딩 벡터(330)를 기초로 잔기 임베딩 벡터(430)가 생성될 수 있다. 예를 들어, 잔기 임베딩 벡터(430)는 다음의 수학식 3을 기초로 생성될 수 있다.

【0066】 **【수학식 3】**

$$I_A = LayerNorm(X_A + E_p + E_c')$$

【0068】여기서, I_A는 잔기 임베딩 벡터(430)를 나타내고, X_A는 아미노산 레벨 임베딩 벡터(330)를 나타낼 수 있다. 또한, $E_p \in R^{L \times d}$는 위치 임베딩 벡터(420)를 나타내고, E_c'는 체인 임베딩 벡터(410)를 나타낼 수 있다. 이 경우, 체인 임베딩 벡터(410)는 각각의 단백질 체인에 대응하는 임베딩 벡터들 $E_c \in R^{L \times d}$이 병합되어 형성될 수 있다. 또한, LayerNorm은 정규화 레이어 및/또는 정규화 연산을 나타낼 수 있다.

【0069】일 실시예에 따르면, 잔기 임베딩 벡터(430)는 인공지능 모델(440)에 제공될 수 있다. 여기서, 인공지능 모델(440)은 합성곱 신경망 기반의 모델일 수 있다. 인공지능 모델(440)은 잔기 임베딩 벡터(430)를 기초로 단백질 서열 상의 인접한 아미노산 잔기 사이의 지역적인 특징을 추출할 수 있다. 이 경우, 인공지능

모델(440)의 출력값으로부터 어텐션 모듈(450)의 입력으로 사용되는 키($K \in R^{L \times d}$) 및 밸류($V \in R^{L \times d}$)가 산출될 수 있다.

【0070】 한편, 단백질 레벨 임베딩 벡터(340)를 기초로 어텐션 모듈(450)의 입력으로 사용되는 쿼리($Q \in R^{L \times d}$)가 산출될 수 있다. 예를 들어, 쿼리는 다음의 수학식 4를 기초로 산출될 수 있다.

【0072】 **【수학식 4】**

$$Q = LayerNorm(I_p + E_p + E_c')$$

【0074】 여기서, I_p는 단백질 레벨 임베딩 벡터(340)를 나타낼 수 있다. 또한, $E_p \in R^{L \times d}$는 위치 임베딩 벡터(420)를 나타내고, E_c'는 체인 임베딩 벡터(410)를 나타낼 수 있다. 또한, LayerNorm은 정규화 레이어 및/또는 정규화 연산을 나타낼 수 있다.

【0075】 일 실시예에 따르면, 어텐션 모듈(450)은 키, 밸류 및 쿼리를 입력받아 단백질에 포함된 아미노산 각각의 위치와 단백질을 구성하는 하위 서열 간의 연관성을 결정할 수 있다. 예를 들어, 어텐션 모듈(450)은 어텐션 레이어(attention layer)에 키, 밸류 및 쿼리를 제공하는 것에 의해 다음의 수학식 5를

기초로 연관성을 결정할 수 있다.

【0077】 **【수학식 5】**

$$Multihead(Q,K,V) = Concat(head_1, \dots, head_H)W^O$$

【0078】 $head_h = Attention(QW_h^Q, KW_h^k, VW_h^V)$

【0079】 $$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

【0081】 여기서, $W_h^Q \in R^{d\times d_q}$, $W_h^K \in R^{d\times d_k}$, $W_h^V \in R^{d\times d_v}$, $W^O \in R^{d\times d_{model}}$ 은 프로젝트 파라미터 행렬(project parameter matrices)을 나타내고, h는 헤드(head)의 수를 나타낼 수 있다.

【0082】 어텐션 레이어의 출력값은 다음의 수학식 6과 같이 완전 연결 피드 포워드 네트워크(fully connected feed-forward network)에 제공될 수 있다. 여기서, 완전 연결 피드 포워드 네트워크는 GELU(Gaussian error linear unit) 활성 함수를 갖는 2개의 선형 변환으로 구성될 수 있다.

【0084】【수학식 6】

$$FFN(x) = \sigma(xW_1 + b_1)W_2 + b_2$$

【0086】이 후, 안정적인 그래디언트(gradient)를 제공하고 학습 속도를 가속화하기 위해 잔차 연결(residual connection) 및 레이어 정규화가 적용될 수 있다. 이에 따라, 단백질을 구성하는 하위 서열 사이의 의존성 정보를 포함하는 어텐션 값($O_A \in R^{L \times d_{model}}$)이 생성될 수 있다.

【0087】어텐션 모듈(450)에 의해 생성된 어텐션 값과 단백질 레벨 임베딩 벡터(340)는 분류기(460)에 제공될 수 있다. 예를 들어, 어텐션 값 및 단백질 레벨 임베딩 벡터(340)는 서로 연결되어 대표 벡터를 형성할 수 있으며, 분류기(460)는 대표 벡터를 제공받아 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정할 수 있다. 분류기(460)는 4개의 은닉층을 포함하는 완전 연결 레이어(fully connected layer)를 포함할 수 있으며, 대표 벡터를 완전 연결 레이어를 통과시키는 것에 의해 단백질 서열의 각각의 위치가 결합 부위인지 여부를 판정할 수 있다.

【0088】일 실시예에 따르면, 분류기(460)는, 가중 교차 엔트로피 손실 함수(weighted cross-entropy loss function)를 기초로 학습될 수 있다. 가중 교차 엔트로피 손실 함수는 다음의 수학식 7과 같이 결정될 수 있다.

【0090】 **【수학식 7】**

$$(p_i, l_i) = -\omega_p \sum_{l_i=1} \log(\sigma(p_i)) - \omega_n \sum_{l_i=0} \log(1 - \sigma(p_i))$$

【0092】 여기서, σ는 시그모이드 함수(sigmoid function)을 나타내고, $\omega_p = \frac{|P|+|N|+1}{P+1}$ 및 $\omega_n = \frac{|P|+|N|+1}{N+1}$은 가중치를 나타낼 수 있다. 이 경우, $|P|$ 및 $|N|$은 각각 포지티브 샘플(positive sample)과 네거티브 샘플(negative sample)의 수를 나타낼 수 있다. 이와 같이, 가중 교차 엔트로피 손실 함수를 이용하여 분류기(460)를 학습시키는 경우, 포지티브 샘플 및 네거티브 샘플 간의 클래스 불균형(class imbalance) 문제가 해소될 수 있다.

【0093】 도 5는 본 개시의 일 실시예에 따른 인공지능 모델(500)의 예시적인 구조를 나타내는 도면이다. 상술된 것과 같이, 인공지능 모델(500)은 합성곱 신경망 기반의 모델일 수 있다. 여기서, 합성곱 신경망은, 복수의 합성곱 신경망 블록(CNN block)을 포함하는 1차원 합성곱 신경망(1D-CNN)일 수 있다.

【0094】 도시된 것과 같이, 합성곱 신경망 기반의 모델을 구성하는 합성곱 신경망 모듈(CNN 모듈)은 계층적 특징을 추출하도록 적층된 3개의 합성곱 신경망 블록을 포함할 수 있다. 이 경우, 적층된 3개의 합성곱 신경망 블록은 서로 다른

확장률(dilation rate)을 가질 수 있다. 또한, 각각의 합성곱 신경망 블록은 서로 다른 커널 폭을 갖는 3개의 CNN 레이어를 포함할 수 있다. 이와 같은 구조를 기초로, 합성곱 신경망 기반의 모델은 다음의 수학식 8과 같이 특징맵을 생성할 수 있다.

【0096】 **【수학식 8】**

$$y_c[i] = \sum_{j=1}^{k} x[i + r \cdot j] w_c[j]$$

【0098】 여기서, r은 확장률을 나타내고, c는 채널(channel)의 수를 나타내며, k는 커널 폭을 나타낼 수 있다. 수학식 8을 이용하여 각각의 합성곱 신경망 블록으로부터 산출된 특징맵들이 결합되어 합성곱 신경망 기반의 모델의 최종적인 출력값을 결정할 수 있다.

【0099】 도 6은 본 개시의 일 실시예에 따른 리간드 결합 부위를 예측하는 모델들 간의 성능을 비교한 그래프(600)이다. 도시된 예에서, 본 개시에 따른 모델은 'Pseq2Sites'이며, 종래 기술은 'HoTS', 'BiRDS', 'Fpocket', 'P2Rank', 'DeepSurf' 및 'DeepPocket' 등의 모델을 포함한다. 모델들의 예측 성능은 'COACH420' 데이터셋으로 평가되었으나, 'HOLO4K', 'NRC-HiQ set1', 'NRC-HiQ

set2' 등의 데이터셋이 이용될 수도 있다.

【0100】모델의 성능은 성공률(success rate)을 기초로 판단될 수 있다. 여기서, 성공률은 다음의 수학식 9와 같이 단백질 서열을 구성하는 전체 포켓(pocket)의 수 대비 모델에 의해 옳게 예측된 포켓의 수를 기초로 산출될 수 있다. 여기서, 포켓은 결합 부위의 집합을 지칭할 수 있다.

【0102】**【수학식 9】**

$$SR(\%) = \frac{no.of\ correctly\ identified\ pockets}{total\ number\ of\ pockets}$$

【0104】모델에 의해 옳게 예측된 것인지 여부는, 다음의 수학식 10과 같이 특정 포켓에 포함된 결합 부위의 수 대비 예측된 결합 부위의 수가 유의 수준(significance level) 이상인지 여부를 기초로 판단될 수 있다.

【0106】**【수학식 10】**

$$\frac{|R_{known} \cap R_{pred}|}{|R_{known}|} > \delta$$

【0108】여기서, R_{known}는 특정 포켓 내에 존재하는 결합 부위의 수이고, R_{pred}는 모델에 의해 예측된 결합 부위의 수를 나타낼 수 있다. 또한, δ는 유의 수준을 나타낼 수 있다.

【0109】도시된 예에서, 유의 수준이 0.5일 때, 본 개시에 따른 'Pseq2Sites' 모델의 성공률은 96.8%로서, 'Fpocket' 모델 대비 2배 이상의 성능을 보이고, 'HoTS' 대비 6배의 성능을 보인 것을 확인할 수 있다. 또한, 유의 수준이 각각 0.9, 0.95 및 0.99인 경우에도 기존 기술 대비 가장 우수한 성능을 보인 것을 쉽게 확인할 수 있다.

【0110】도 7은 본 개시의 일 실시예에 따른 리간드 결합 부위를 예측하는 모델들 간의 예측 결과를 시각화한 이미지(700)이다. 이미지(700) 상의 (a) 도면 및 (c) 도면은 본 개시에 따른 'Pseq2Sites' 모델의 예측 결과를 나타내고, (b) 도면 및 (d) 도면은 종래 기술 중 하나인 'P2Rank' 모델의 예측 결과를 나타낸다. 여기서, 'COA', 'LBT' 및 'CSF'는 단백질 서열에 포함된 결합 부위를 나타낸다.

【0111】도시된 예에서, 각각의 결합 부위는 단백질 서열 상에서 대응하는 색상(빨간색 및 파란색)으로 표시되고, 각각의 모델에 의해 예측된 결과는 회색의 하이라이트(highlight)로 표시될 수 있다. (a) 도면과 (b) 도면을 참조하면, 'Pseq2Sites' 모델은 'R'에 대응하는 하나의 결합 부위만을 예측하지 못한 반면, 'P2Rank' 모델은 'S', 'D', 'K', 'N' 및 'T'의 5개의 결합 부위를 예측하지 못한 것을 간단히 확인할 수 있다. 또한, (c) 도면과 (d) 도면을 참조하면,

'Pseq2Sites' 모델은 모든 결합 부위를 예측한 반면, 'P2Rank' 모델은 'R', 'G', 'T' 및 'I'의 4개의 결합 부위를 예측하지 못한 것을 확인할 수 있어, 'Pseq2Sites' 모델이 'P2Rank' 모델이 비해 성능이 우수한 것을 시각적으로 확인할 수 있다.

【0112】도 8은 본 개시의 일 실시예에 따른 단백질 서열 기반의 리간드 결합 부위 예측 방법(800)의 예시를 나타내는 흐름도이다. 단백질 서열 기반의 리간드 결합 부위 예측 방법(800)은 프로세서(예를 들어, 컴퓨팅 장치의 적어도 하나의 프로세서)에 의해 수행될 수 있다. 단백질 서열 기반의 리간드 결합 부위 예측 방법(800)은 프로세서가 단백질 서열을 입력받음으로써 개시될 수 있다(S810).

【0113】프로세서는 입력된 단백질 서열을 기초로 아미노산 레벨 임베딩 벡터 및 단백질 레벨 임베딩 벡터를 생성할 수 있다(S820). 예를 들어, 프로세서는 입력된 단백질 서열을 구성하는 각각의 아미노산에 대한 임베딩을 수행하여 아미노산 레벨 임베딩 벡터를 생성하고, 아미노산 레벨 임베딩 벡터의 평균값을 기초로 단백질 레벨 임베딩 벡터를 생성할 수 있다.

【0114】프로세서는 생성된 아미노산 레벨 임베딩 벡터 및 단백질 레벨 임베딩 벡터를 이용하여 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정할 수 있다(S830). 일 실시예에 따르면, 프로세서는 아미노산 레벨 임베딩 벡터를 이용하여 잔기 임베딩 벡터를 생성하고, 생성된 잔기 임베딩 벡터를 학습된 인공지능 모델에 제공하여 인접한 아미노산 잔기 사이의 지역적 특징을 나타내는 키 및 밸류를 산출할 수 있다. 또한, 프로세서는 단백질 레벨 임베딩 벡터를

이용하여 쿼리를 산출할 수 있다.

【0115】 잔기 임베딩 벡터 생성을 위해, 프로세서는 단백질 서열을 구성하는 단백질 체인과 연관된 체인 임베딩 벡터를 생성하고, 단백질 서열을 구성하는 각각의 아미노산의 위치와 연관된 위치 임베딩 벡터를 생성할 수 있다. 그리고 나서, 프로세서는 생성된 체인 임베딩 벡터 및 위치 임베딩 벡터를 아미노산 레벨 임베딩 벡터에 더하여 잔기 임베딩 벡터를 생성할 수 있다. 추가적으로, 프로세서는 체인 임베딩 벡터 및 위치 임베딩 벡터를 단백질 레벨 임베딩 벡터에 더하여 쿼리를 산출할 수 있다.

【0116】 그리고 나서, 프로세서는 산출된 키, 밸류 및 쿼리를 어텐션 모듈에 제공하여 어텐션 값을 산출하고, 산출된 어텐션 값을 기초로 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정할 수 있다. 예를 들어, 프로세서는 어텐션 값 및 단백질 레벨 임베딩 벡터를 연결하여 단백질 서열에 대응하는 대표 벡터를 생성하고, 생성된 대표 벡터를 결합 부위 여부를 판정하도록 학습된 분류기에 제공하여 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정할 수 있다.

【0117】 도 9는 본 개시의 일 실시예에 따른 컴퓨팅 장치(100)의 하드웨어 구성을 나타내는 블록도이다. 컴퓨팅 장치(100)는 메모리(910), 프로세서(920), 통신 모듈(930) 및 입출력 인터페이스(940)를 포함할 수 있으며, 도 9에 도시된 바와 같이, 컴퓨팅 장치(100)는 통신 모듈(930)을 이용하여 네트워크를 통해 정보 및/또는 데이터를 통신할 수 있도록 구성될 수 있다.

【0118】메모리(910)는 비-일시적인 임의의 컴퓨터 판독 가능한 기록매체를 포함할 수 있다. 일 실시예에 따르면, 메모리(910)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(100)에 포함될 수 있다. 또한, 메모리(910)에는 운영체제와 적어도 하나의 프로그램 코드가 저장될 수 있다.

【0119】이러한 소프트웨어 구성요소들은 메모리(910)와는 별도의 컴퓨터에서 판독 가능한 기록매체로부터 로딩될 수 있다. 이러한 별도의 컴퓨터에서 판독 가능한 기록매체는 이러한 컴퓨팅 장치(100)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(930)을 통해 메모리(910)에 로딩될 수도 있다. 예를 들어, 적어도 하나의 프로그램은 개발자들 또는 애플리케이션의 설치 파일을 배포하는 파일 배포 시스템이 통신 모듈(930)을 통해 제공하는 파일들에 의해 설치되는 컴퓨터 프로그램에 기반하여 메모리(910)에 로딩될 수 있다.

【0120】프로세서(920)는 기본적인 산술, 로직 및 입출력 연산을 수행함으로써, 컴퓨터 프로그램의 명령을 처리하도록 구성될 수 있다. 명령은 메모리(910) 또

는 통신 모듈(930)에 의해 다른 사용자 단말(미도시) 또는 다른 외부 시스템으로 제공될 수 있다.

【0121】 통신 모듈(930)은 네트워크를 통해 사용자 단말(미도시)과 컴퓨팅 장치(100)가 서로 통신하기 위한 구성 또는 기능을 제공할 수 있으며, 컴퓨팅 장치(100)가 외부 시스템(일례로 별도의 클라우드 시스템 등)과 통신하기 위한 구성 또는 기능을 제공할 수 있다. 일례로, 컴퓨팅 장치(100)의 프로세서(920)의 제어에 따라 제공되는 제어 신호, 명령, 데이터 등이 통신 모듈(930)과 네트워크를 거쳐 사용자 단말 및/또는 외부 시스템의 통신 모듈을 통해 사용자 단말 및/또는 외부 시스템으로 전송될 수 있다.

【0122】 또한, 컴퓨팅 장치(100)의 입출력 인터페이스(940)는 컴퓨팅 장치(100)와 연결되거나 컴퓨팅 장치(100)가 포함할 수 있는 입력 또는 출력을 위한 장치(미도시)와의 인터페이스를 위한 수단일 수 있다. 도 9에서는 입출력 인터페이스(940)가 프로세서(920)와 별도로 구성된 요소로서 도시되었으나, 이에 한정되지 않으며, 입출력 인터페이스(940)가 프로세서(920)에 포함되도록 구성될 수 있다. 컴퓨팅 장치(100)는 도 9의 구성요소들보다 더 많은 구성요소들을 포함할 수 있다. 그러나, 대부분의 종래기술적 구성요소들을 명확하게 도시할 필요성은 없다.

【0123】 컴퓨팅 장치(100)의 프로세서(920)는 복수의 사용자 단말 및/또는 복수의 외부 시스템으로부터 수신된 정보 및/또는 데이터를 관리, 처리 및/또는 저장하도록 구성될 수 있다.

【0124】 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 언어를 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.

【0125】상술된 방법 및/또는 다양한 실시예들은, 입력 데이터를 기초로 동작하거나 출력 데이터를 생성함으로써, 임의의 기능, 함수 등을 처리, 저장 및/또는 관리하는 하나 이상의 컴퓨터 프로그램을 실행하도록 구성된 하나 이상의 프로세서에 의해 수행될 수 있다. 예를 들어, 본 개시의 방법 및/또는 다양한 실시예는 FPGA(Field Programmable Gate Array) 또는 ASIC(Application Specific Integrated Circuit)과 같은 특수 목적 논리 회로에 의해 수행될 수 있으며, 본 개시의 방법 및/또는 실시예들을 수행하기 위한 장치 및/또는 시스템은 FPGA 또는 ASIC와 같은 특수 목적 논리 회로로서 구현될 수 있다.

【0126】컴퓨터 프로그램을 실행하는 하나 이상의 프로세서는, 범용 목적 또는 특수 목적의 마이크로 프로세서 및/또는 임의의 종류의 디지털 컴퓨팅 장치의

하나 이상의 프로세서를 포함할 수 있다. 프로세서는 읽기 전용 메모리, 랜덤 액세스 메모리의 각각으로부터 명령 및/또는 데이터를 수신하거나, 읽기 전용 메모리와 랜덤 액세스 메모리로부터 명령 및/또는 데이터를 수신할 수 있다. 본 발명에서, 방법 및/또는 실시예들을 수행하는 컴퓨팅 장치의 구성 요소들은 명령어들을 실행하기 위한 하나 이상의 프로세서, 명령어들 및/또는 데이터를 저장하기 위한 하나 이상의 메모리 디바이스를 포함할 수 있다.

【0127】일 실시예에 따르면, 컴퓨팅 장치는 데이터를 저장하기 위한 하나 이상의 대용량 저장 장치와 데이터를 주고받을 수 있다. 예를 들어, 컴퓨팅 장치는 자기 디스크(magnetic disc) 또는 광 디스크(optical disc)로부터 데이터를 수신하거나/수신하고, 자기 디스크 또는 광 디스크로 데이터를 전송할 수 있다. 컴퓨터 프로그램과 연관된 명령어들 및/또는 데이터를 저장하기에 적합한 컴퓨터 판독 가능한 저장 매체는, EPROM(Erasable Programmable Read-Only Memory), EEPROM(Electrically Erasable PROM), 플래시 메모리 장치 등의 반도체 메모리 장치를 포함하는 임의의 형태의 비 휘발성 메모리를 포함할 수 있으나, 이에 한정되지 않는다. 예를 들어, 컴퓨터 판독 가능한 저장 매체는 내부 하드 디스크 또는 이동식 디스크와 같은 자기 디스크, 광 자기 디스크, CD-ROM 및 DVD-ROM 디스크를 포함할 수 있다.

【0128】사용자와의 상호 작용을 제공하기 위해, 컴퓨팅 장치는 정보를 사용자에게 제공하거나 디스플레이하기 위한 디스플레이 장치(예를 들어, CRT (Cathode Ray Tube), LCD(Liquid Crystal Display) 등) 및 사용자가 컴퓨팅 장치 상에 입력

및/또는 명령 등을 제공할 수 있는 포인팅 장치(예를 들어, 키보드, 마우스, 트랙볼 등)를 포함할 수 있으나, 이에 한정되지 않는다. 즉, 컴퓨팅 장치는 사용자와의 상호 작용을 제공하기 위한 임의의 다른 종류의 장치들을 더 포함할 수 있다. 예를 들어, 컴퓨팅 장치는 사용자와의 상호 작용을 위해, 시각적 피드백, 청각 피드백 및/또는 촉각 피드백 등을 포함하는 임의의 형태의 감각 피드백을 사용자에게 제공할 수 있다. 이에 대해, 사용자는 시각, 음성, 동작 등의 다양한 제스처를 통해 컴퓨팅 장치로 입력을 제공할 수 있다.

【0129】본 발명에서, 다양한 실시예들은 백엔드 구성 요소(예: 데이터 서버), 미들웨어 구성 요소(예: 애플리케이션 서버) 및/또는 프론트 엔드 구성 요소를 포함하는 컴퓨팅 시스템에서 구현될 수 있다. 이 경우, 구성 요소들은 통신 네트워크와 같은 디지털 데이터 통신의 임의의 형태 또는 매체에 의해 상호 연결될 수 있다. 예를 들어, 통신 네트워크는 LAN(Local Area Network), WAN(Wide Area Network) 등을 포함할 수 있다.

【0130】본 명세서에서 기술된 예시적인 실시예들에 기반한 컴퓨팅 장치는, 사용자 디바이스, 사용자 인터페이스(UI) 디바이스, 사용자 단말 또는 클라이언트 디바이스를 포함하여 사용자와 상호 작용하도록 구성된 하드웨어 및/또는 소프트웨어를 사용하여 구현될 수 있다. 예를 들어, 컴퓨팅 장치는 랩톱(laptop) 컴퓨터와 같은 휴대용 컴퓨팅 장치를 포함할 수 있다. 추가적으로 또는 대안적으로, 컴퓨팅 장치는, PDA(Personal Digital Assistants), 태블릿 PC, 게임 콘솔(game console), 웨어러블 디바이스(wearable device), IoT(internet of things) 디바이스,

VR(virtual reality) 디바이스, AR(augmented reality) 디바이스 등을 포함할 수 있으나, 이에 한정되지 않는다. 컴퓨팅 장치는 사용자와 상호 작용하도록 구성된 다른 유형의 장치를 더 포함할 수 있다. 또한, 컴퓨팅 장치는 이동 통신 네트워크 등의 네트워크를 통한 무선 통신에 적합한 휴대용 통신 디바이스(예를 들어, 이동 전화, 스마트 전화, 무선 셀룰러 전화 등) 등을 포함할 수 있다. 컴퓨팅 장치는, 무선 주파수(RF; Radio Frequency), 마이크로파 주파수(MWF; Microwave Frequency) 및/또는 적외선 주파수(IRF; Infrared Ray Frequency)와 같은 무선 통신 기술들 및/또는 프로토콜들을 사용하여 네트워크 서버와 무선으로 통신하도록 구성될 수 있다.

【0131】본 발명에서 특정 구조적 및 기능적 세부 사항을 포함하는 다양한 실시예들은 예시적인 것이다. 따라서, 본 개시의 실시예들은 상술된 것으로 한정되지 않으며, 여러 가지 다른 형태로 구현될 수 있다. 또한, 본 발명에서 사용된 용어는 일부 실시예를 설명하기 위한 것이며 실시예를 제한하는 것으로 해석되지 않는다. 예를 들어, 단수형 단어 및 상기는 문맥상 달리 명확하게 나타내지 않는 한 복수형도 포함하는 것으로 해석될 수 있다.

【0132】본 발명에서, 달리 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함하여 본 명세서에서 사용되는 모든 용어는 이러한 개념이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 갖는다. 또한, 사전에 정의된 용어와 같이 일반적으로 사용되는 용어들은 관련 기술의 맥락에서의 의미와 일치하는 의미를 갖는 것으로 해석되어야 한다.

【0133】본 명세서에서는 본 발명이 일부 실시예들과 관련하여 설명되었지만, 본 개시의 발명이 속하는 기술분야의 통상의 기술자가 이해할 수 있는 본 개시의 범위를 벗어나지 않는 범위에서 다양한 변형 및 변경이 이루어질 수 있다. 또한, 그러한 변형 및 변경은 본 명세서에 첨부된 특허청구의 범위 내에 속하는 것으로 생각되어야 한다.

【부호의 설명】

【0135】100: 컴퓨팅 장치

102: 단백질 서열

104: 결합 부위

110: 임베딩 모듈

120: 인공지능 모델

130: 어텐션 모듈

140: 분류기

【청구범위】

【청구항 1】

적어도 하나의 프로세서에 의해 수행되는 단백질 서열 기반의 리간드(ligand) 결합 부위 예측 방법으로서,

단백질 서열을 입력받는 단계;

상기 입력된 단백질 서열을 기초로 아미노산 레벨 임베딩 벡터(amino acid-level embedding vector) 및 단백질 레벨 임베딩 벡터(protein-level embedding vector)를 생성하는 단계; 및

상기 생성된 아미노산 레벨 임베딩 벡터 및 상기 단백질 레벨 임베딩 벡터를 이용하여 상기 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정하는 단계;

를 포함하는 단백질 서열 기반의 리간드 결합 부위 예측 방법.

【청구항 2】

제1항에 있어서,

상기 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정하는 단계는,

상기 아미노산 레벨 임베딩 벡터를 이용하여 잔기 임베딩 벡터(residue embedding vector)를 생성하는 단계;

상기 생성된 잔기 임베딩 벡터를 학습된 인공지능 모델에 제공하여 인접한 아미노산 잔기 사이의 지역적 특징을 나타내는 키(key) 및 밸류(value)를 산출하는 단계;

상기 단백질 레벨 임베딩 벡터를 이용하여 쿼리(query)를 산출하는 단계;

상기 산출된 키, 밸류 및 쿼리를 어텐션 모듈(attention module)에 제공하여 어텐션 값을 산출하는 단계; 및

상기 산출된 어텐션 값을 기초로 상기 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정하는 단계;

를 포함하는 단백질 서열 기반의 리간드 결합 부위 예측 방법.

【청구항 3】

제2항에 있어서,

상기 아미노산 레벨 임베딩 벡터를 이용하여 잔기 임베딩 벡터를 생성하는 단계는,

상기 단백질 서열을 구성하는 단백질 체인과 연관된 체인 임베딩 벡터(chain embedding vector)를 생성하는 단계;

상기 단백질 서열을 구성하는 각각의 아미노산의 위치와 연관된 위치 임베딩 벡터(position embedding vector)를 생성하는 단계; 및

상기 생성된 체인 임베딩 벡터 및 상기 위치 임베딩 벡터를 상기 아미노산

레벨 임베딩 벡터에 더하여 상기 잔기 임베딩 벡터를 생성하는 단계;

를 포함하는 단백질 서열 기반의 리간드 결합 부위 예측 방법.

【청구항 4】

제2항에 있어서,

상기 단백질 레벨 임베딩 벡터를 이용하여 쿼리를 산출하는 단계는,

상기 단백질 서열을 구성하는 단백질 체인과 연관된 체인 임베딩 벡터를 생성하는 단계;

상기 단백질 서열을 구성하는 각각의 아미노산의 위치와 연관된 위치 임베딩 벡터를 생성하는 단계; 및

상기 생성된 체인 임베딩 벡터 및 상기 위치 임베딩 벡터를 상기 단백질 레벨 임베딩 벡터에 더하여 상기 쿼리를 산출하는 단계;

를 포함하는 단백질 서열 기반의 리간드 결합 부위 예측 방법.

【청구항 5】

제2항에 있어서,

상기 산출된 어텐션 값을 기초로 상기 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정하는 단계는,

상기 어텐션 값 및 상기 단백질 레벨 임베딩 벡터를 연결하여 상기 단백질

서열에 대응하는 대표 벡터를 생성하는 단계; 및

상기 생성된 대표 벡터를 결합 부위 여부를 판정하도록 학습된 분류기(classifier)에 제공하여 상기 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정하는 단계;

를 포함하는 단백질 서열 기반의 리간드 결합 부위 예측 방법.

【청구항 6】

제5항에 있어서,

상기 분류기는, 4개의 은닉층(hidden layer)을 갖는 완전 연결 레이어(fully connected layer)를 포함하는, 단백질 서열 기반의 리간드 결합 부위 예측 방법.

【청구항 7】

제5항에 있어서,

상기 분류기는, 가중 교차 엔트로피 손실 함수(weighted cross-entropy loss function)를 기초로 학습되는, 단백질 서열 기반의 리간드 결합 부위 예측 방법.

【청구항 8】

제2항에 있어서,

상기 인공지능 모델은, 합성곱 신경망(convolutional neural network; CNN)

기반의 모델을 포함하는, 단백질 서열 기반의 리간드 결합 부위 예측 방법.

【청구항 9】

제8항에 있어서,

상기 합성곱 신경망은, 복수의 합성곱 신경망 블록(CNN block)을 포함하는 1차원 합성곱 신경망(1D-CNN)인, 단백질 서열 기반의 리간드 결합 부위 예측 방법.

【청구항 10】

제9항에 있어서,

상기 복수의 합성곱 신경망 블록에 포함된 각각의 합성곱 신경망 블록은 서로 다른 커널 폭(kernel width)을 갖는 3개의 합성곱 신경망 레이어(CNN layer)를 포함하는, 단백질 서열 기반의 리간드 결합 부위 예측 방법.

【청구항 11】

제1항에 있어서,

상기 입력된 단백질 서열을 기초로 아미노산 레벨 임베딩 벡터 및 단백질 레벨 임베딩 벡터를 생성하는 단계는,

상기 입력된 단백질 서열을 구성하는 각각의 아미노산에 대한 임베딩을 수행하여 아미노산 레벨 임베딩 벡터를 생성하는 단계; 및

상기 생성된 아미노산 레벨 임베딩 벡터의 평균값을 기초로 단백질 레벨 임베딩 벡터를 생성하는 단계;

를 포함하는 단백질 서열 기반의 리간드 결합 부위 예측 방법.

【청구항 12】

제1항 내지 제11항 중 어느 한 항에 따른 방법을 컴퓨터에서 실행하기 위해 컴퓨터 판독 가능한 기록 매체에 저장된 컴퓨터 프로그램.

【청구항 13】

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,

단백질 서열을 입력받고,

상기 입력된 단백질 서열을 기초로 아미노산 레벨 임베딩 벡터 및 단백질 레벨 임베딩 벡터를 생성하고,

상기 생성된 아미노산 레벨 임베딩 벡터 및 상기 단백질 레벨 임베딩 벡터를 이용하여 상기 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정하기 위한 명령어들을 포함하는 컴퓨팅 장치.

【요약서】

【요약】

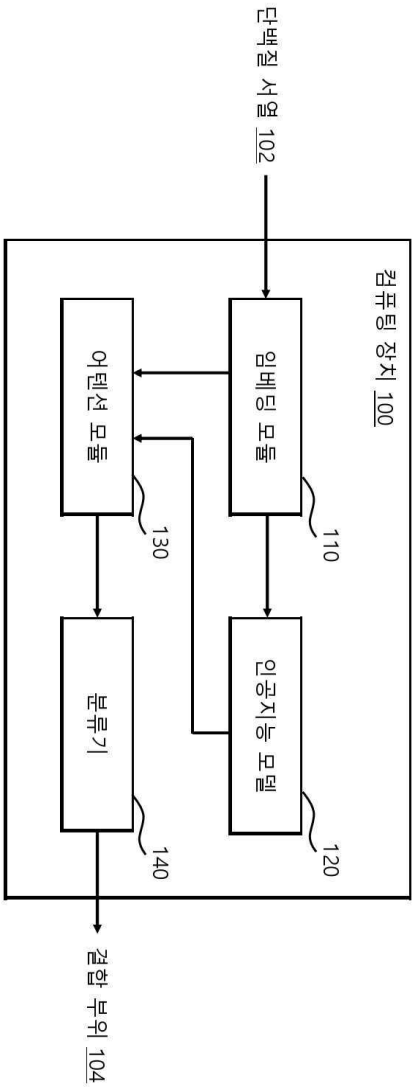
본 개시는 단백질 서열 기반의 리간드 결합 부위 예측 방법에 관한 것이다. 단백질 서열 기반의 리간드 결합 부위 예측 방법은, 단백질 서열을 입력받는 단계, 입력된 단백질 서열을 기초로 아미노산 레벨 임베딩 벡터 및 단백질 레벨 임베딩 벡터를 생성하는 단계 및 생성된 아미노산 레벨 임베딩 벡터 및 단백질 레벨 임베딩 벡터를 이용하여 단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정하는 단계를 포함한다.

【대표도】

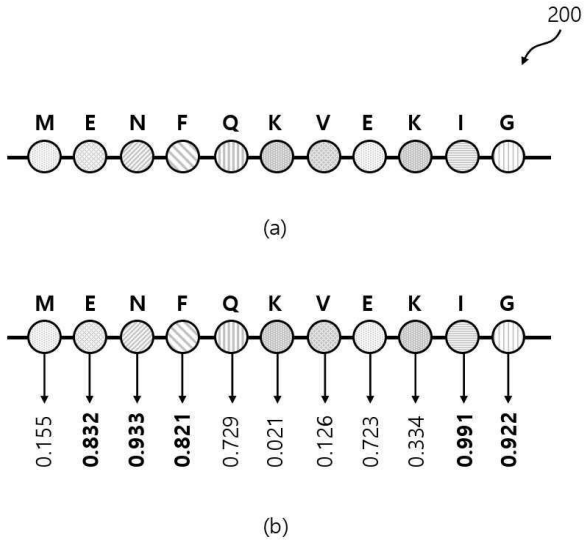
도 1

【도면】

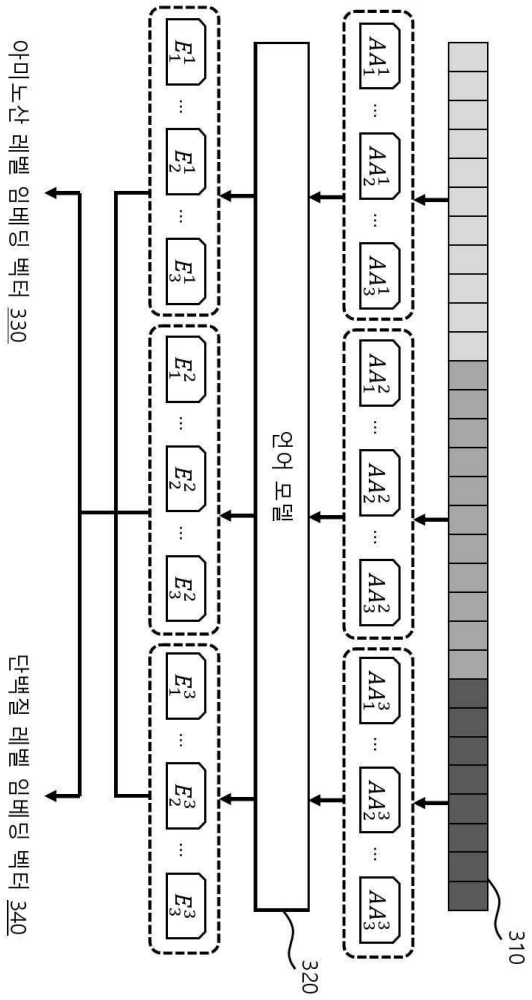
【도 1】


단백질 서열 102
컴퓨팅 장치 100
임베딩 모듈
어텐션 모듈
분류기
인공지능 모델
110
120
130
140
결합 부위 104

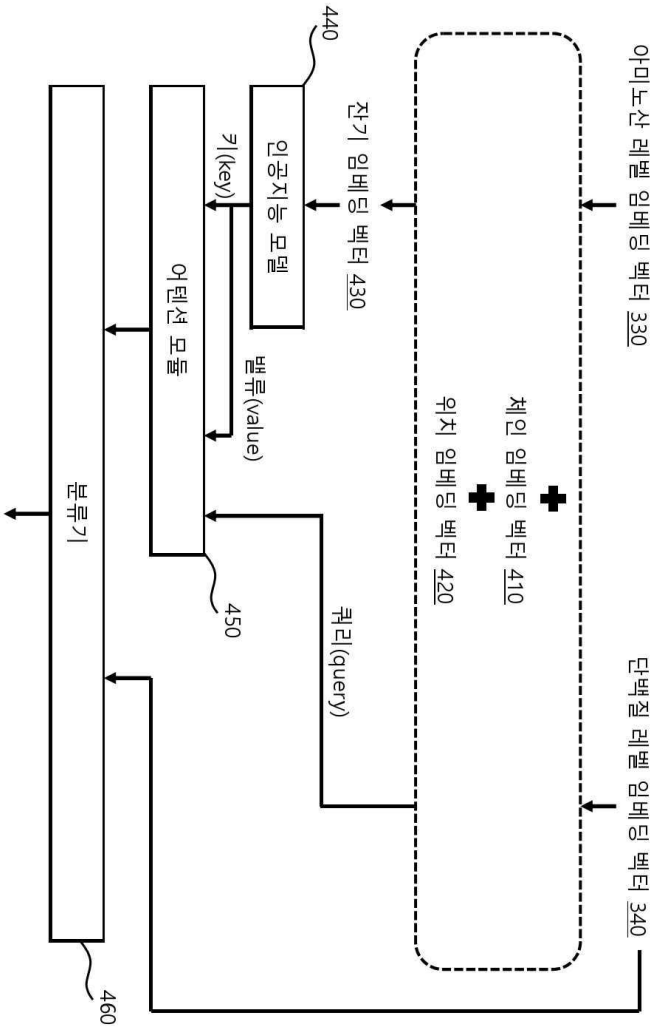
【도 2】


200
M E N F Q K V E K I G
(a)
M E N F Q K V E K I G
0.155
0.832
0.933
0.821
0.729
0.021
0.126
0.723
0.334
0.991
0.922
(b)

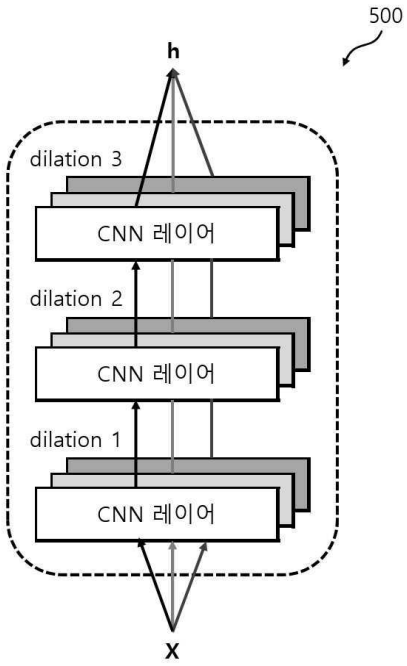
【도 3】



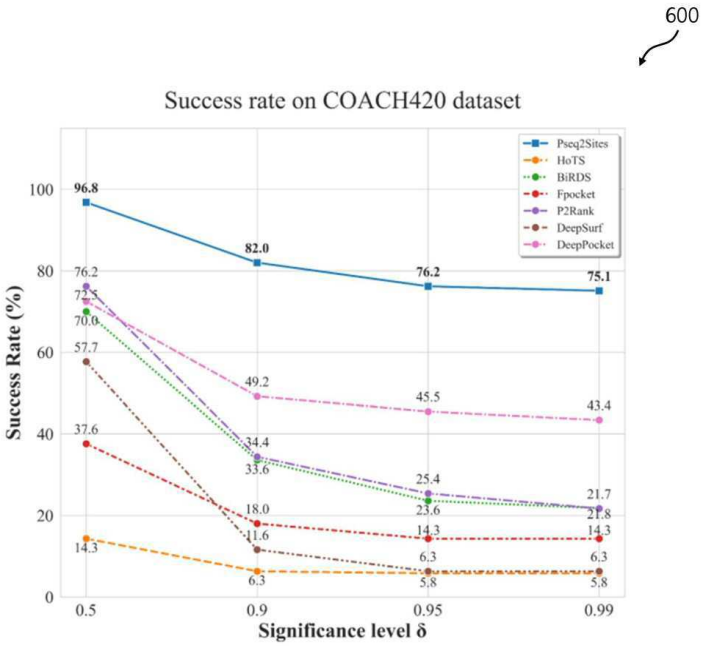
【도 4】


아미노산 레벨 임베딩 벡터 330
체인 임베딩 벡터 410
위치 임베딩 벡터 420
단백질 레벨 임베딩 벡터 340
잔기 임베딩 벡터 430
440
인공지능 모델
키(key)
밸류(value)
어텐션 모듈
450
쿼리(query)
분류기
460

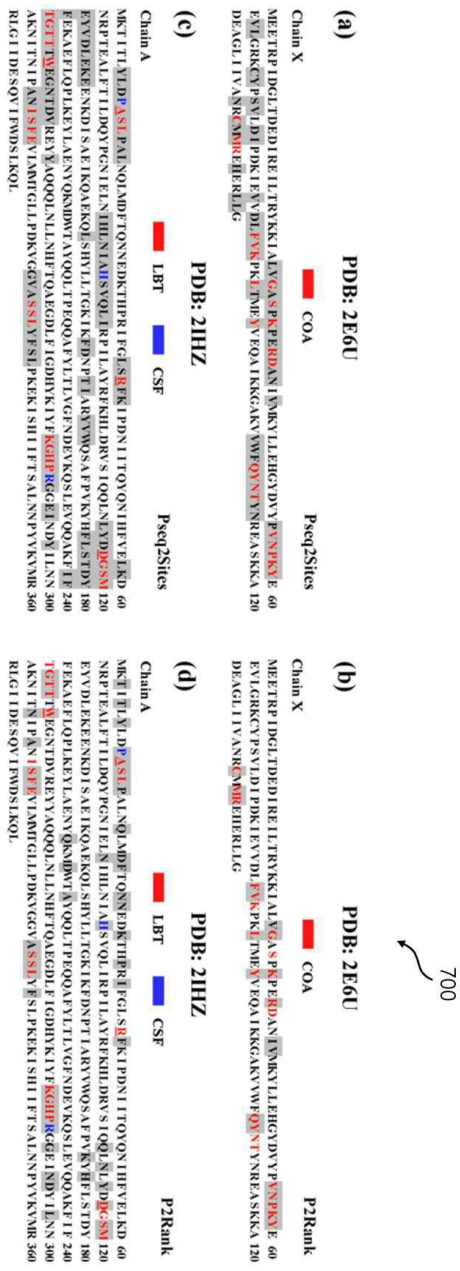
【도 5】


500
h
dilation 3
CNN 레이어
dilation 2
CNN 레이어
dilation 1
CNN 레이어
X

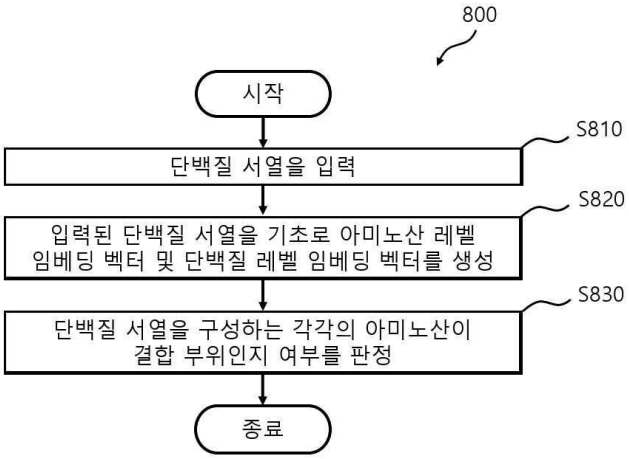
【도 6】


600
Success rate on COACH420 dataset
Pseq2Sites
HoTS
BiRDS
Fpocket
P2Rank
DeepSurf
DeepPocket
Success Rate (%)
Significance level δ
0.5
0.9
0.95
0.99
96.8
82.0
76.2
75.1

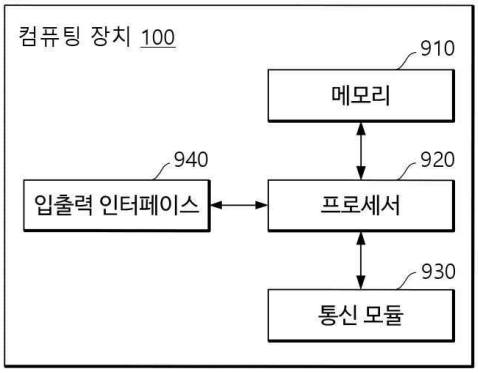
【도 7】


(a)
PDB: 2E6U
COA
Pseq2Sites
Chain X
(c)
PDB: 2IHZ
LBT
CSF
Pseq2Sites
Chain A
(b)
PDB: 2E6U
COA
P2Rank
Chain X
(d)
PDB: 2IHZ
LBT
CSF
P2Rank
Chain A
700

【도 8】


800
시작
S810
단백질 서열을 입력
S820
입력된 단백질 서열을 기초로 아미노산 레벨 임베딩 벡터 및 단백질 레벨 임베딩 벡터를 생성
S830
단백질 서열을 구성하는 각각의 아미노산이 결합 부위인지 여부를 판정
종료

【도 9】


컴퓨팅 장치 100
910
메모리
940
920
입출력 인터페이스
프로세서
930
통신 모듈