(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2024년11월14일
(11) 등록번호 10-2730477
(24) 등록일자 2024년11월11일

(51) 국제특허분류(Int. Cl.)
*G16B 40/20* (2019.01) *G06N 20/00* (2019.01)
*G06N 3/04* (2023.01) *G06N 3/08* (2023.01)
*G16B 15/30* (2019.01) *G16B 30/00* (2019.01)
*G16B 50/00* (2019.01)
(52) CPC특허분류
*G16B 40/20* (2019.02)
*G06N 20/00* (2021.08)
(21) 출원번호 10-2023-0101278
(22) 출원일자 2023년08월02일
심사청구일자 2023년08월02일
(56) 선행기술조사문헌
A. Robert 외, "Impact of target interaction on small-molecule drug disposition: an overlooked area", Nature Reviews, 2018.*
M. A. Thafar 외, "Affinity2Vec: drug-target binding affinity prediction through representation learning, graph mining, and machine learning", Scientific Reports, 12:4751, 2022.*
Y. Wang 외, "In silico prediction of human intravenous Pharmacokinetic parameters with improved accuracy", JCIM 59, pp.3968-3980, 2019.08.12.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
**충남대학교 산학협력단**
대전광역시 유성구 대학로 99 (궁동, 충남대학교)
(72) 발명자
**김영국**
대전광역시 유성구 엑스포로 448 엑스포아파트 402동 1004호
**윤휘열**
대전광역시 유성구 죽동로 321, 104동 2103호
*(뒷면에 계속)*
(74) 대리인
**이은철, 김재문**

전체 청구항 수 : 총 6 항 심사관 : 성경아

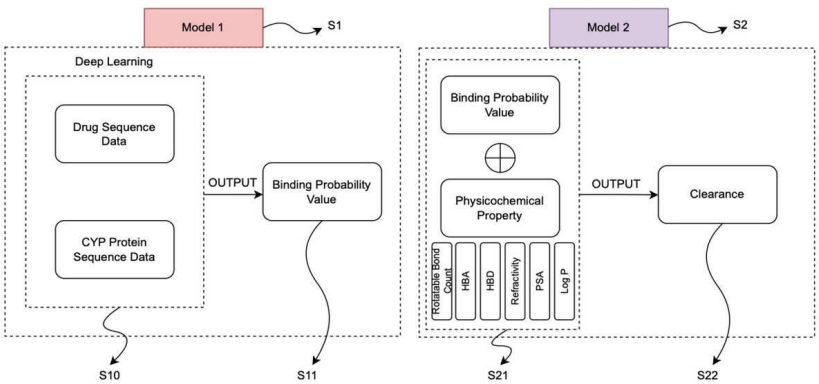
(54) 발명의 명칭 **인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템**

## (57) 요 약

본 발명은 약물과 표적 단백질의 결합 확률을 활용하여 체내에 약물이 오래 머물거나 짧게 머무는 것을 알려 주는 가장 중요한 지표인 청소율 예측시스템에 관한 것이다. 이를 위하여 본 발명은 수집된 약물과 타겟 단백질의 시퀀스 데이터를 서브워드 토크나이저를 사용해 단어집합을 생성하는 토크나이징 단계, 학습을 위해 시퀀스 데이

(뒷면에 계속)

*대 표 도* - 도1



터를 임베딩 해주는 임베딩 단계, 길이가 서로 다른 약물과 표적 단백질의 시퀀스 데이터 길이를 통일시켜주는 데이터 길이 통일 단계, 인공신경망을 통해 학습하여 약물과 표적 단백질의 결합 확률을 예측하는 결합 확률 예측 단계, 약물 데이터에서 도출된 6개 이상의 추가 입력값들을 추가하여 통계 기반 모델에 학습하여 청소율을 예측하는 청소율 예측 단계로 구성되어 청소율을 예측하여, 후보 약물과 단백질 사이의 상호작용을 가지고 청소율을 구할 수 있는 효과가 있다.

(52) CPC특허분류
*G06N 3/04* (2023.01)
*G06N 3/08* (2023.01)
*G16B 15/30* (2019.02)
*G16B 30/00* (2019.02)
*G16B 50/00* (2019.02)

(72) 발명자
**하정민**
대전광역시 동구 용운로 203, 111동 1305호(
용운동, e편한세상대전에코포레아파트)
**이현정**
대전광역시 중구 대종로 713, 1동 904호

**손유지**
대전광역시 서구 한밭대로 656번길 22, 3층

이 발명을 지원한 국가연구개발사업

| | |
|---|---|
| 과제고유번호 | 1711179294 |
| 과제번호 | 00155857 |
| 부처명 | 과학기술정보통신부 |
| 과제관리(전문)기관명 | 정보통신기획평가원 |
| 연구사업명 | 인공지능융합혁신인재양성 |
| 연구과제명 | 인공지능융합혁신인재양성(충남대학교) |
| 기 여 율 | 1/1 |
| 과제수행기관명 | 충남대학교산학협력단 |
| 연구기간 | 2023.01.01 ~ 2023.12.31 |

# 명 세 서

## 청구범위

### 청구항 1

약물 시퀀스 데이터와 단백질 시퀀스 데이터로부터 단어의 일부 또는 조합으로 서브워드 단위로 분리하여 단어 집합을 생성하는 토크나이징부,

상기 토크나이징부에서 만들어진 단어 집합을 사용하여 학습모델에 학습시키기 위해 시퀀스 데이터를 임베딩하는 임베딩부,

길이가 서로 다른 약물 시퀀스 데이터와 표적 단백질 시퀀스 데이터의 길이를 통일시켜 전처리하는 패딩부,

전처리된 데이터를 학습모델을 통해 학습하여 약물과 표적 단백질의 결합 확률을 추출하는 결합확률 예측부 및

약물데이터에서 도출된 추가 입력값들을 추가하여 투여하는 약물과 그 약물의 표적이 되는 단백질의 상호작용을 통해 구해진 결합 확률값을 가지고 학습모델을 사용하여 청소율을 구하기 위한 청소율 예측부를 포함하되,

상기 청소율 예측부는 약물 데이터에서 도출된 추가 입력값(Rotatable Bond Count, HBA, HBD, Refractiviry, PSA, Log p)을 결합확률 예측부에서 추출된 결합 확률에 추가하여 통계 기반 학습모델에 학습하여 청소율을 예측하는 것을 특징으로 하는 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템.

### 청구항 2

제1항에 있어서,

상기 토크나이징부는 약물 시퀀스 데이터와 단백질 시퀀스 데이터를 서브워드 토크나이저를 사용하여 단어 집합을 생성하되, 토크나이징된 단어들은 각각 고유한 정수 인덱스로 변환하는 것을 특징으로 하는 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템.

### 청구항 3

제1항에 있어서,

상기 임베딩부는 토크나이징부에서 생성된 단어 집합을 사용하여 학습 모델에 학습시키기 위해 시퀀스 데이터를 임베딩하되, 토크나이징부를 통해 각각 정수 인덱스를 해당 단어의 임베딩 벡터로 변환하는 것을 특징으로 하는 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템.

### 청구항 4

제1항에 있어서,

상기 패딩부는 패딩 토큰을 활용하여 길이가 서로 다른 약물 시퀀스 데이터와 표적 단백질 시퀀스 데이터의 길이를 통일시켜 전처리하되, 이 길이는 데이터셋에서 가장 긴 시퀀스의 길이를 기준으로 선택하여 통일시킬 길이를 결정하는 것을 특징으로 하는 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템.

### 청구항 5

제1항에 있어서,

상기 청소율 예측부는

상기 결합확률 예측부의 결과값과 약물 데이터로부터 도출되는 6개의 추가 입력값(Rotatable Bond Count, HBA, HBD, Refractiviry, PSA, Log p)을 가지고 통계 기반의 머신러닝 모델을 통해 청소율을 예측하는 것을 특징으로 하는 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템.

### 청구항 6

인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템을 이용한 방법에 있어서,

(a)상기 청소율 예측 시스템이 약물 시퀀스 데이터와 단백질 시퀀스 데이터로부터 단어의 일부 또는 조합으로 서브워드 단위로 분리하여 단어 집합을 생성하는 단계,

(b)상기 청소율 예측 시스템이 상기 (a)단계에서 만들어진 단어 집합을 사용하여 학습모델에 학습시키기 위해 시퀀스 데이터를 임베딩하는 단계,

(c)상기 청소율 예측 시스템이 상기 (b)단계의 임베딩 결과, 패딩 토큰을 활용하여 길이가 서로 다른 약물 시퀀스 데이터와 표적 단백질 시퀀스 데이터의 길이를 통일시켜 전처리하는 단계,

(d)상기 청소율 예측 시스템이 전처리가 완료된 데이터를 학습모델을 통해 학습하여 약물과 표적 단백질의 결합 확률을 예측하는 단계,

(e)상기 청소율 예측 시스템이 약물데이터에서 도출된 추가 입력값들을 추가하여 투여하는 약물과 그 약물의 표적이 되는 단백질의 상호작용을 통해 구해진 결합 확률값을 가지고 학습모델을 사용하여 청소율을 예측하는 단계를 포함하되,

상기 청소율을 예측하는 단계는 약물 데이터에서 도출된 추가 입력값(Rotatable Bond Count, HBA, HBD, Refractiviry, PSA, Log p)을 결합확률 예측부에서 추출된 결합 확률에 추가하여 통계 기반 학습모델에 학습하여 청소율을 예측하는 것을 특징으로 하는 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측방법.

## 발명의 설명

### 기 술 분 야

[0001] 본 발명은 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템에 관한 것으로, 더욱 상세하게는 약물과 표적 단백질의 결합 확률을 활용하여 청소율을 예측하기 위한 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템에 관한 것이다.

### 배 경 기 술

[0002] Nature Review Drug Discovery에 따르면 전임상부터 최종 상용화까지 평균 성공률은 9.6%이다. 신약개발에서의 1단계인 발견단계가 전체 소요기간에서 절반 정도를 차지하고 소요비용이 대략 19억 정도로 후보 물질은 5천 개에서 1만 개의 후보 약물이 발견된다. 반면에 2단계인 전임상과 임상에서 필요한 비용이 앞서 단계의 필요한 비용보다 큰 폭으로 증가하는데 전임상 약 38억 원으로 최소 10개~250개까지 약물이 줄어들게 된다. 임상 1상에서 3상을 거쳐 시판 전까지 최소 약 350억 원 등이 소요비용으로 쓰이나, 시판 가능한 약물은 1개가 안 되는 예도 있다.

[0003] 따라서 현재의 신약개발에 있어서 신약개발 연구원과 제약사들은 이러한 비용 및 시간을 줄이는 방법에 연구하고 있다. 앞서 이야기한 것과 같이 후보 약물을 찾고, 관련 단백질에 대한 결합 확률을 예측하는 방법들은 이미 인공지능 기술을 통해 많은 연구가 진행되어있다. 그러나, 신약 개발에 있어서 기존의 표적 단백질과 후보 약물의 결합 확률을 통해 예측하는 것뿐만 아니라 약물이 몸 안으로 흡수되고, 시간에 따라 혈액과 각 조직에서 얼마나 분포되어 약효를 발효하는지를 의미하는 약동학적인 결과를 알려주는 것도 매우 중요하다.

[0004] 기존에는 이 약동학적인 결과를 예측하는 연구에 대해서 많이 진행되어있지 않았다. 약동학 파라미터의 예로는 청소율, 소실 반감기, 최고혈장농도 등이 있는데 그중에서 청소율(Clearance;CL)은 약물이 체내에 오래 머물거나 짧게 머무는 것을 알려주는 전임상 실험에 있어서 단백질과 약물의 결합 확률과 같이 상당히 중요한 지표이다.

[0005] 청소율은 약물이 전신에 도달하기 전에 대사과정을 거치게 되는데 대사란, 섭취한 약물이 생체 내에서 흡수되어 일어나는 화학적 작용으로 주로 간에서 이루어지는 것으로 이때의 간에서 대사를 하게 도움을 주는 단백질 촉매제가 Cytochrome P450 (CYP450)이다. 대사과정을 통하여 약물은 몸 밖으로 배출되게 되는데 이때의 배설기관에서의 물질을 함유하는 혈장의 양을 의미한다. 이 청소율을 기존에는 표적 단백질을 고려하지 않고 후보 약물의 특성만을 고려하여 예측해왔다.

[0006] 예를 들어, 기존 연구에서 약동학적인 결과를 알려주는 모델 중 MolDesigner라는 이름의 데모 프로그램이 있다.

이것은 코로나 후보 약물 15개와 관련 표적 단백질을 가지고 COVID-19을 목표로 만들어진 모델로써 후보 약물과 표적 단백질 사이의 연관 관계에 대하여 중점적으로 발표하였고 그 성능을 인정받았다. 하지만 MolDesigner의 경우 청소율 예측 성능에 있어서 표적 단백질을 고려하지 않고 약물의 청소율 값만을 고려하여 좋지 않은 예측 성능을 보였다. 따라서 기존의 청소율을 예측하는데에 있어 한계가 있음을 알 수 있다.

## 선행기술문헌

### 특허문헌

[0007] (특허문헌 0001) 공개특허 제10-2022-0015479호(2022.02.08. 공개)

## 발명의 내용

### 해결하려는 과제

[0008] 본 발명은 상술한 문제를 해결하고자, 투여하는 약물과 그 약물의 표적이 되는 단백질의 상호작용을 통해 구해진 결합 확률값을 가지고 인공지능을 사용하여 청소율을 구하기 위한 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템을 제공함에 있다.

[0009] 또한 본 발명의 다른 목적은 후보 약물과 단백질 사이의 상호작용을 가지고 청소율을 구하는 것으로, 약물 시퀀스 데이터와 표적 단백질의 시퀀스 데이터의 결합 확률을 인공 신경망 모델을 통해 출력하고, 출력된 결과값과 약물 데이터로부터 도출되는 6개의 추가 입력값(Rotatable Bond Count, HBA, HBD, Refractiviry, PSA, Log p)을 가지고 통계 기반의 머신러닝 모델을 통해 청소율을 예측하는 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템을 제공함에도 있다.

### 과제의 해결 수단

[0010] 본 발명은 수집된 약물 시퀀스 데이터와 단백질 시퀀스 데이터를 서브워드 토크나이저를 사용해 단어 집합을 생성하는 토크나이징부, 상기 토크나이징부에서 만들어진 단어 집합을 사용하여 학습모델에 학습시키기 위해 시퀀스 데이터를 임베딩하는 임베딩부, 길이가 서로 다른 약물 시퀀스 데이터와 표적 단백질 시퀀스 데이터의 길이를 통일시켜 전처리하는 패딩부, 전처리된 데이터를 인공신경망을 통해 학습하여 약물과 표적 단백질의 결합 확률을 추출하는 결합확률 예측부 및 약물데이터에서 도출된 추가 입력값들을 추가하여 투여하는 약물과 그 약물의 표적이 되는 단백질의 상호작용을 통해 구해진 결합 확률값을 가지고 인공지능을 사용하여 청소율을 구하기 위한 청소율 예측부를 포함한다.

[0011] 한편, 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템을 이용한 방법에 있어서, (a)상기 청소율 예측 시스템이 수집된 약물 시퀀스 데이터와 단백질 시퀀스 데이터를 서브워드 토크나이저를 사용해 단어 집합을 생성하는 단계, (b)상기 청소율 예측 시스템이 상기 (a)단계에서 만들어진 단어 집합을 사용하여 학습모델에 학습시키기 위해 시퀀스 데이터를 임베딩하는 단계, (c)상기 청소율 예측 시스템이 상기 (b)단계의 임베딩 결과, 패딩 토큰을 활용하여 길이가 서로 다른 약물 시퀀스 데이터와 표적 단백질 시퀀스 데이터의 길이를 통일시켜 전처리하는 단계, (d)상기 청소율 예측 시스템이 전처리가 완료된 데이터를 인공신경망을 통해 학습하여 약물과 표적 단백질의 결합확률을 예측하는 단계, (e)상기 청소율 예측 시스템이 약물데이터에서 도출된 추가 입력값들을 추가하여 투여하는 약물과 그 약물의 표적이 되는 단백질의 상호작용을 통해 구해진 결합 확률값을 가지고 인공지능을 사용하여 청소율을 예측하는 단계를 포함한다.

### 발명의 효과

[0012] 본 발명에 따르면, 글로벌 신약개발 시, 평균 1조 원의 개발 비용과 10년의 개발 기간이 소요되며 성공률이 1/5,000으로 매우 낮다는 하이리스크(high-risk)를 가지고 있지만 개발 성공시 장기간 고수익이 날 수 있는 하이리턴(high-return)을 보여준다.

[0013] 이상과 같이 본 발명에 따르면 이러한 비용적, 시간적 문제의 발생 가능성을 크게 줄여줄 수 있을 것으로 보인다. 또한 과학 기술적 기대효과로서 신약개발의 방향설정에 대한 의사결정에 긍정적 영향을 끼칠 수 있으며 신약개발의 성공률도 높일 수 있을 것으로 사료된다.

## 도면의 간단한 설명

[0014] 도 1은 본 발명의 일 실시예에 따른 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템에서, 약물과 단백질 시퀀스의 입력에 따라 청소율을 예측하는 방법을 설명하기 위한 도면이다.

도 2는 본 발명의 일 실시예에 따른 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템의 구성도이다.

도 3은 본 발명의 일 실시예에 따른 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템에서, 약물과 단백질 시퀀스의 입력에 따라 청소율을 예측하는 방법을 순서대로 도시한 플로우 차트이다.

도 4는 본 발명의 일 실시예에 따른 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템의 결과값 도출에 사용된 Model1의 내부 구조를 나타낸 도면이다.

도 5는 본 발명의 일 실시예에 따른 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 방법을 나타낸 흐름도이다.

## 발명을 실시하기 위한 구체적인 내용

[0015] 본 발명은 약물과 표적 단백질의 결합 확률을 활용하여 체내에 약물이 오래 머물거나 짧게 머무는 것을 알려 주는 가장 중요한 지표인 청소율 예측시스템에 관한 것이다.

[0016] 본 발명의 일 실시예에 따른 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템은 신약개발에 있어 중요한 지표인 청소율을 예측하는데 있어, 표적 단백질 데이터를 추가적으로 사용하고 이를 인공 신경망을 통해 더 정확한 예측을 수행하도록 한다.

[0017] 본 실시예에 따르면, 약물의 대사와 관련된 단백질인 CYP450을 가지고 학습을 하여 신약개발 1단계에서 걸러진 5천 개~1만 개의 후보 약물을 전부 직접 실험해보지 않고, 약동학적 예측결과인 청소율을 보여주는 약물과 체내 노출을 정량화할 필요가 있다. 이는 약물들을 빠르게 순위를 매겨, 의사결정에 도움을 주고, 개발 시간을 배분하면서 효과적인 약물을 먼저 추려 실제 실험에 들어가는 시간과 비용 절감에 중요한 역할을 할 수 있을 것이다.

[0018] 본 발명의 일 실시예에 따른 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템에 있어서, 가장 큰 목표는 약물과 표적 단백질 사이의 상호작용을 가지고 청소율을 구하는 것이다.

[0019] 후보 약물과 표적 단백질인 CYP450간의 상호작용을 위한 구조정보를 이용하는데 약물의 구조정보는 SMILES로 복잡한 화합물을 한 줄로 표현할 수 있는 형식으로 사용한다.

[0020] 단백질 구조정보는 단백질 구조를 나타내는 형식인 아미노산 서열(Amino acid sequence)를 사용하며 두 구조정보를 나타내는 데이터의 길이가 일정하지 않음에 따라 자연어 처리 FCS mining을 사용한다.

[0021] 본 발명을 첨부된 도면을 참조하여 상세히 설명하면 다음과 같다.

[0022] 도 1은 본 발명의 일 실시예에 따른 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템에서, 약물과 단백질 시퀀스의 입력에 따라 청소율을 예측하는 방법을 설명하기 위한 도면이다.

[0023] 도 1은 크게 Model1과 Model2로 나누어진 형태이고, Model1은 S1이고 Model2는 S2를 나타낸다.

[0024] S10은 약물과 타겟 단백질 시퀀스를 입력으로 넣어주는 부분이다. S11은 S10을 통해 입력값으로 받은 두 개의 시퀀스 데이터를 인공신경망을 통해 추론하여 결합 확률을 예측하여 결과 값을 출력하는 부분이다.

[0025] S21은 출력된 결과값과 약물 데이터로부터 도출되는 6개 이상의 추가 입력값(Rotatable Bond Count, HBA, HBD, Refractiviry, PSA, Log p) 등을 입력으로 통계학적 모델에 입력해주는 부분이다. 마지막으로 S22는 최종 추론을 통하여 청소율을 출력하는 부분이다.

[0026] 본 발명의 일 실시예에 따른 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템에서, 약물과 단백질 시퀀스의 입력에 따라 청소율을 예측하는 Model1과 Model2에 대해 부연설명하면, 다음과 같다.

[0027] "Model1"은 약물과 단백질 시퀀스의 입력에 따라 결합 확률을 예측하는 부분을 담당하는 모델이다. 도 1에서 "Model1"은 S1로 표현되며, S10과 S11을 포함한다.

[0028] S10은 약물과 타겟 단백질 시퀀스를 입력으로 받아들이는 부분이다. 이 부분은 약물 데이터와 표적 단백질 시퀀스 데이터를 전처리하고 인공신경망 모델에 입력으로 넣어준다.

[0029] S11은 S10을 통해 입력받은 두 개의 시퀀스 데이터를 인공신경망을 통해 추론하여 결합 확률을 예측하고, 예측 결과 값을 출력하는 부분이다. 이렇게 예측된 결합 확률은 약물과 표적 단백질 간의 상호작용 가능성을 나타내며, Model1의 주요 출력으로 사용된다.

[0030] "Model2"는 Model1에서 추론된 결합 확률과 약물 데이터로부터 도출되는 추가 입력값을 활용하여 최종적으로 청소율을 예측하는 부분을 담당하는 모델이다. 도 1에서 "Model2"는 S2로 표현되며, S21과 S22를 포함한다.

[0031] S21은 Model1에서 출력된 결합 확률과 약물 데이터로부터 도출된 6개 이상의 추가 입력값(Rotatable Bond Count, HBA, HBD, Refractiviry, PSA, Log p) 등을 입력으로 받는다. 이 입력값들과 결합 확률을 통계학적 모델에 입력으로 주어진다.

[0032] S22는 S21을 통해 최종적으로 추론된 청소율 값을 출력하는 부분이다. Model2의 주요 출력으로 사용되며, 약물-표적 단백질 상호작용 예측 시스템에서 최종적인 청소율 예측 결과를 제공한다.

[0033] 이렇게 "Model1"과 "Model2"가 각각 다른 역할을 수행하여 약물-표적 단백질 상호작용 모델을 구성하며, 최종적으로 약물의 청소율을 예측하는데 기여한다.

[0034] 도 2는 본 발명의 일 실시예에 따른 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템의 구성도이다.

[0035] 도 2에 도시된 바와 같이, 본 발명의 일 실시예에 따른 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템(10)은 토크나이징부(110), 임베딩부(120), 패딩부(130), 결합확률 예측부(140), 청소율 예측부(150)를 포함한다.

[0036] 토크나이징부(110)는 청소율 예측에 있어서, 수집된 약물 시퀀스 데이터와 단백질 시퀀스 데이터를 서브워드 토크나이저를 사용해 단어 집합을 생성하기 위한 구성이다.

[0037] 이러한 토크나이징부(110)는 약물 시퀀스 데이터와 단백질 시퀀스 데이터를 서브워드 토크나이저를 사용하여 단어 집합을 생성한다. 예를 들어, 약물 시퀀스 데이터가 "O=C(O)CCCC[C@@H]1CCSS1"이고, 단백질 시퀀스 데이터가 "MALIPDLAMETWLLLA"이라면, 서브워드 토크나이저는 이러한 문장들을 서브워드(단어의 일부 혹은 조합) 단위로 분리하여 단어 집합을 만들 수 있다. 이렇게 토크나이징된 단어들은 고유한 정수 인덱스로 매핑되어 임베딩과 다른 처리에 사용된다.

[0038] 서브워드 토크나이저는 문장을 단어 단위로만 분리하는 것이 아니라, 단어를 더 작은 단위로 분리하거나 조합하는 방식을 사용하여 단어 집합을 형성한다. 이렇게 하면, 긴 단어를 여러 개의 작은 단위로 분해하여 처리하고, 더 적은 단어로 표현될 수 있는 효과를 얻을 수 있다. 이렇게 생성된 단어들은 고유한 정수 인덱스로 매핑되어 컴퓨터가 처리하기 쉬운 형태로 변환된다.

[0039] 예를 들어, 토크나이징부를 사용하여 주어진 약물 시퀀스 데이터와 단백질 시퀀스 데이터를 토크나이즈하면 다음과 같은 결과를 얻을 수 있다.

[0040] 약물 시퀀스 데이터: "O=C(O)CCCC[C@@H]1C"

[0041] 단백질 시퀀스 데이터: "MALIPDLAMETWLLLA"

[0042] 토크나이저를 사용하여 이러한 문장들을 서브워드(단어의 일부 혹은 조합) 단위로 분리하면 다음과 같은 결과를 얻을 수 있습니다.

[0043] 약물 시퀀스 데이터 토크나이즈 결과:

[0044] ["O","=","C(O)","C","C","C","C","[C@@H]1","C"]

[0045] 단백질 시퀀스 데이터 토크나이즈 결과:

[0046] ["M","A","L","I","P","D","L","A","M","E"]

[0047] 이렇게 토크나이징된 단어들은 각각 고유한 정수 인덱스로 매핑되어 학습 모델에 입력으로 사용되며, 이를 통해 약물-표적 단백질 상호작용을 예측하는 인공신경망 모델을 효과적으로 학습할 수 있다.

[0048] 임베딩부(120)는 토크나이징부(110)에서 만들어진 단어 집합을 사용하여 학습모델에 학습시키기 위해 시퀀스 데이터를 임베딩한다.

[0049] 이러한 임베딩부(120)는 토크나이징부에서 생성된 단어 집합을 사용하여 학습 모델에 학습시키기 위해 시퀀스 데이터를 임베딩하는 구성이다. 예를 들어, "Aspirin"과 "kinase"라는 단어들은 토크나이징부를 통해 각각 정수 인덱스 1과 2로 매핑된다. 임베딩은 이러한 정수 인덱스를 해당 단어의 임베딩 벡터로 변환한다. 예를 들어, 인덱스 1에 해당하는 "Aspirin"의 임베딩 벡터는 [0.2, 0.5, -0.1]이 될 수 있다. 이렇게 변환된 임베딩 벡터들은 학습 모델의 입력으로 사용된다.

[0050] 보다 구체적으로 설명하면, 토크나이징부(110)에서 만들어진 단어 집합은 서브워드 토크나이저를 사용하여 약물 시퀀스 데이터와 표적 단백질 시퀀스 데이터를 처리한 결과이다. 이 단어 집합은 각각의 단어나 토큰들을 고유한 정수 인덱스로 매핑한 것이며, 이제 이 단어 집합을 사용하여 학습모델에 학습시키기 위해 시퀀스 데이터를 임베딩하는 과정을 설명하면 다음과 같다.

[0051] 예를 들어, 토크나이징부를 거쳐 약물 시퀀스 데이터가 "O=C(O)CCCC[C@@H]1C"로 토큰화되었고, 해당 토큰들은 각각 다음과 같이 매핑되었다고 가정하면,

[0052] "O": 1

[0053] "=": 2

[0054] "C(O)": 3

[0055] "C": 4

[0056] "C": 4

[0057] "C": 4

[0058] "C": 4

[0059] "[C@@H]1": 5

[0060] "C": 4

[0061] 마찬가지로, 표적 단백질 시퀀스 데이터가 "MALIPDLAMETWLLLA"로 토큰화되었고, 해당 토큰들은 다음과 같이 매핑되었다고 가정하면,

[0062] "M": 5

[0063] "A": 6

[0064] "L": 7

[0065] "I": 8

[0066] "P": 9

[0067] "D": 10

[0068] "L": 7

[0069] "A": 6

[0070] 이제 이러한 단어 집합을 사용하여 학습모델에 학습시키기 위해 시퀀스 데이터를 임베딩해야 한다. 임베딩은 토크나이징된 각 토큰들을 벡터 형태로 변환하는 과정이다. 예를 들어, 각 토큰들은 위에서 정의한 단어 집합의 인덱스와 일치한다. 따라서 "O"은 인덱스 1, "="는 인덱스 2, "C(O)"는 인덱스 3 등과 같이 표현된다.

[0071] 임베딩은 이러한 정수 인덱스를 해당 단어의 임베딩 벡터로 변환한다. 각 단어의 임베딩 벡터는 학습 모델의 입력으로 사용된다. 임베딩은 일반적으로 미리 학습된 워드 임베딩(Word Embedding) 모델을 사용하거나, 학습 데이터를 통해 모델을 함께 학습시키는 방식으로 이루어진다.

[0072] 예를 들어, 임베딩 벡터의 차원을 50차원으로 설정했다면, 각 토큰들은 50차원의 실수 벡터로 표현된다. 예를 들어 "Aspirin"의 임베딩 벡터는 [0.2, 0.5, -0.1, ..., 0.3]과 같이 표현될 수 있다. 이러한 임베딩 벡터들은

약물 시퀀스와 표적 단백질 시퀀스의 모든 토큰에 대해 생성된다.

[0073] 이렇게 생성된 임베딩 벡터들은 학습 모델의 입력으로 사용된다. 이후, 임베딩된 약물 시퀀스와 표적 단백질 시퀀스는 결합확률 예측부(140)와 같은 모델에 입력으로 들어가며, 인공신경망을 통해 학습하고 결합확률을 예측하는데 활용된다.

[0074] 패딩부(130)는 패딩 토큰을 활용하여 길이가 서로 다른 약물 시퀀스 데이터와 표적 단백질 시퀀스 데이터의 길이를 통일시켜 전처리하는 기능을 수행한다.

[0075] 본 실시예에 따른 패딩부(130)는 길이가 서로 다른 약물 시퀀스 데이터와 표적 단백질 시퀀스 데이터의 길이를 통일시키는 기능을 수행하는 구성이다. 패딩은 데이터의 길이를 맞추는 과정으로, 더 짧은 시퀀스에는 특정 값을 채워서 시퀀스의 길이를 늘리는 방법을 사용한다. 이를 통해 모든 시퀀스 데이터가 동일한 길이를 가지게 된다.

[0076] 예를 들어, 토크나이징된 약물 시퀀스 데이터가 다음과 같이 표현되었다고 가정하면,

[0077] "O=C(O)CCCC[C@@H]1C" → [1, 2, 3, 4, 4, 4, 4, 5, 4]

[0078] 표적 단백질 시퀀스 데이터가 다음과 같이 표현되었다고 가정한다.

[0079] "MALIPDLAMETWLLLA" → [5, 6, 7, 8, 9, 10, 7, 6]

[0080] 위의 두 시퀀스 데이터는 서로 다른 길이를 가지고 있고, 패딩부는 이러한 시퀀스들을 동일한 길이로 만들어야 한다. 패딩을 적용하기 위해서는 먼저 통일시킬 길이를 결정해야 한다. 이 길이는 데이터셋에서 가장 긴 시퀀스의 길이를 기준으로 선택한다. 예를 들어, 약물 시퀀스와 표적 단백질 시퀀스 중에서 가장 긴 시퀀스의 길이가 10이라고 가정한다.

[0081] 패딩 앞 방식으로 한다면, 가장 긴 시퀀스의 길이가 10이므로, 모든 시퀀스 데이터의 길이를 10으로 맞추기 위해 부족한 길이만큼 패딩 토큰을 앞에 추가한다. 패딩 토큰은 보통 정수 0으로 사용된다.

[0082] 예를 들어, 패딩을 적용한 약물 시퀀스 데이터는 다음과 같이 표현될 수 있다.

[0083] [0, 0, 0, 0, 0, 1, 2, 3, 4, 5] → 길이 10

[0084] 패딩을 적용한 표적 단백질 시퀀스 데이터는 다음과 같이 표현될 수 있다.

[0085] [0, 10, 11, 12, 13, 14, 15, 16, 17, 0] → 길이 10

[0086] 패딩 뒤 방식으로 한다면, 위에서 언급한 앞 방식과 반대로 부족한 길이만큼 패딩 토큰을 뒤에 추가한다. 예를 들어, 패딩을 적용한 약물 시퀀스 데이터는 다음과 같이 표현될 수 있다.

[0087] [1, 2, 3, 4, 5, 6, 7, 8, 9, 0] → 길이 10

[0088] 패딩을 적용한 표적 단백질 시퀀스 데이터는 다음과 같이 표현될 수 있다.

[0089] [10, 11, 12, 13, 14, 15, 16, 17, 0, 0] → 길이 10

[0090] 이렇게 패딩부를 통해 모든 약물 시퀀스와 표적 단백질 시퀀스의 길이가 통일되면, 이 데이터를 학습모델에 입력으로 사용할 수 있다. 패딩으로 인해 시퀀스 중 빈 부분은 의미가 없지만, 이후 임베딩과 인공신경망에서 패딩 토큰의 영향이 없도록 처리되어 진행된다. 이렇게 모든 데이터가 동일한 길이로 맞추어지면, 인공신경망 모델이 더욱 효율적으로 학습되고 예측이 수행될 수 있다.

[0091] 결합확률 예측부(140)는 전처리가 완료된 데이터를 인공신경망을 통해 학습하여 약물과 표적 단백질의 결합 확률을 추출한다. 이러한 결합확률 예측부(140)는 인공신경망을 사용하여 약물과 표적 단백질의 결합 확률을 예측하는 구성으로, 다음과 같이 이루어진다.

[0092] 결합확률 예측부(140)는 패딩부(130)를 거쳐서 길이가 통일된 약물 시퀀스 데이터와 표적 단백질 시퀀스 데이터를 입력데이터로 한다. 이러한 입력데이터는 임베딩된 벡터 형태로 표현되며, 이 데이터를 인공신경망 모델에 입력으로 제공한다.

[0093] 본 실시예에 따른 인공신경망 모델은 결합확률 예측을 위해 딥러닝 모델 중 하나인 다층 퍼셉트론(MLP, Multi-Layer Perceptron)이나 재귀 신경망(RNN, Recurrent Neural Network), 컨볼루션 신경망(CNN, Convolutional Neural Network) 등의 모델을 사용할 수 있고, 이러한 모델은 복잡한 비선형 관계를 학습하여 약물과 표적 단백

질의 결합확률을 예측하는데 사용된다.

[0094] 본 실시예에서는 다층 퍼셉트론(MLP)을 사용하여 결합확률 예측을 설명하면 다음과 같다.

[0095] 먼저, 측정된 약물-단백질 상호작용 데이터를 수집한다. 이 데이터는 다양한 약물과 표적 단백질의 시퀀스 정보와 해당 결합확률(실제값)을 포함한다.

[0096] 이후, 수집된 약물 시퀀스와 표적 단백질 시퀀스 데이터를 토크나이징하여 단어 또는 토큰들로 분리하고, 이를 벡터형태로 임베딩한다. 이 임베딩은 단어 집합을 기반으로 각 토큰을 벡터로 변환하는 과정을 거친다.

[0097] 이후, 다층 퍼셉트론(MLP), 재귀 신경망(RNN), 컨볼루션 신경망(CNN) 등 중 하나의 딥러닝 모델을 선택하며, 다층 퍼셉트론(MLP)으로 설명하기로 한다.

[0098] 약물 시퀀스와 표적 단백질 시퀀스의 데이터를 학습 데이터로 사용한다. 각각의 시퀀스 데이터는 임베딩된 벡터로 표현된다. 학습 데이터에는 각 약물과 표적 단백질의 시퀀스 데이터와 그에 해당하는 결합확률(실제값)이 포함된다.

[0099] 이후, 다층 퍼셉트론(MLP)을 학습시킨다. 학습 데이터를 이용하여 모델이 약물과 표적 단백질의 시퀀스 정보를 학습하고, 복잡한 비선형 관계를 파악한다. 모델은 입력 데이터와 실제 결합확률 사이의 오차를 최소화하는 방향으로 가중치와 편향을 조정하여 학습된다.

[0100] 학습이 완료된 MLP를 사용하여 새로운 약물과 표적 단백질 시퀀스 데이터를 입력으로 받아 결합확률을 예측한다. 입력 데이터는 임베딩을 거친 벡터 형태로 모델에 입력되며, 모델은 입력 데이터를 처리하여 예측된 결합확률을 출력한다.

[0101] 이후, 예측된 결합확률과 실제 결합확률을 비교하여 모델의 성능을 평가한다. 평균 제곱 오차(Mean Squared Error, MSE)나 교차 엔트로피 오차(Cross-Entropy Loss) 등의 손실 함수를 사용하여 예측 값과 실제 값의 오차를 계산하고, 이를 최소화하는 방향으로 모델을 최적화한다. 이렇게 약물과 표적 단백질의 시퀀스 데이터를 입력으로 받아 인공신경망 기반의 모델이 복잡한 비선형 관계를 학습하여 결합확률을 예측하게 된다.

[0102] 청소율 예측부(150)는 약물 데이터에서 도출된 6개 이상의 추가 입력값(Rotatable Bond Count, HBA, HBD, Refractiviry, PSA, Log p)을 결합확률 예측부에서 추출된 결합 확률에 추가하여 통계 기반 모델에 학습하여 청소율을 예측한다.

[0103] 참고로, Rotatable Bond Count (회전 가능한 결합 수)는 약물 분자 내에 회전이 가능한 결합의 수를 나타내는 값으로, 분자의 유연성과 구조적 특성을 나타낸다. 이 값이 높을수록 분자가 더 유연하며, 약물의 화학적 특성과 생체 내 활동성에 영향을 미칠 수 있다.

[0104] 본 실시예에 따른 청소율 예측부(150)의 Rotatable Bond Count (회전 가능한 결합 수) 추가 입력값 도출과 관련하여 부연 설명하면, SMILES 형식 변환과 관련하여, 약물 데이터에서 Rotatable Bond Count를 도출하기 위해서는 해당 약물의 분자 구조를 분석하여 회전 가능한 결합의 수를 계산해야 한다. 이를 수행하기 위해 화학 정보 도구나 라이브러리를 사용할 수 있다. 아래는 예시적인 절차는 다음과 같다.

[0105] 우선 SMILES(Simplified Molecular Input Line Entry System) 형식 변환과 관련하여, 약물 데이터로 주어진 분자 구조를 SMILES 형식으로 변환한다. SMILES는 분자 구조를 간단하게 표현하는 문자열 표기 방법으로, 이를 활용하여 분자 구조를 분석한다. 분자 구조 읽기와 관련하여, SMILES로 표현된 분자 구조를 화학 정보 도구나 라이브러리를 사용하여 읽는다. 대표적으로 RDKit (Cheminformatics Software) 등이 사용될 수 있다. 회전 가능한 결합 확인과 관련하여, 읽어온 분자 구조를 기반으로 각 원자들 사이의 결합 정보를 분석한다. 회전 가능한 결합은 단일 결합(C-C 또는 C-N 등)으로 이루어진 2차원 결합이며, 이러한 결합은 분자의 유연성과 3D 구조를 나타내는데 중요한 역할을 한다. Rotatable Bond Count 계산과 관련하여, 회전 가능한 결합의 수를 계산한다. 각 회전 가능한 결합은 분자 구조에서 회전을 통해 다른 구조로 변화시킬 수 있는 결합을 의미한다. 결합 정보를 분석하여 회전 가능한 결합의 수를 세는 방식으로 Rotatable Bond Count를 도출한다. 예를 들어, 약물로서 아세트아미노펜의 SMILES 표현은 "CC(=O)Nc1ccc(cc1)O"이다. 이를 RDKit와 같은 화학 정보 도구로 읽어온 후 회전 가능한 결합을 계산하면, 아세트아미노펜 분자 내에 회전 가능한 결합이 4개라는 결과를 얻을 수 있다. 따라서 이 경우 Rotatable Bond Count는 4가 된다. 이렇게 계산된 Rotatable Bond Count는 약물의 유연성과 구조적 특징을 나타내는데 사용되며, 약물 데이터에서 도출된 6개 이상의 추가 입력값 중 하나로 도출할 수 있다.

[0106] HBA(수소결합 수용체, Hydrogen Bond Acceptor)는 약물 분자가 수소 결합을 형성할 수 있는 수용체의 수를 의미

한다. 수소 결합은 약물 분자가 단백질과 상호작용하는데 중요한 역할을 할 수 있으며, 약물의 생체 내 활동성과 특성을 예측하는데 사용된다.

[0107] 본 실시예에 따른 청소율 예측부(150)의 HBA(수소결합 수용체, Hydrogen Bond Acceptor) 추가 입력값 도출과 관련하여 부연 설명하면, 약물 데이터에서 HBA (수소결합 수용체)의 입력값을 도출하기 위해서 해당 약물의 분자 구조를 분석하여 수소 수용자가 되는 기능성 원자의 수를 계산해야 한다. 이를 수행하기 위해 화학 정보 도구나 라이브러리를 사용할 수 있다. 아래는 예시적인 절차로 다음과 같다.

[0108] 위에서 언급한 바와 같이, 약물 데이터로 주어진 분자 구조를 SMILES 형식으로 변환하고, SMILES로 표현된 분자 구조를 화학 정보 도구나 라이브러리를 사용하여 읽는다. 대표적으로 RDKit (Cheminformatics Software) 등이 사용될 수 있다. 수소 수용자 확인과 관련하여, 읽어온 분자 구조를 기반으로 각 원자들의 수소 결합을 분석한다. 수소 결합은 원자들 간에 수소 원자가 다른 원자의 비공유 전자쌍과 결합하여 형성되는 결합을 의미한다. 이러한 수소 결합은 약물과 표적 단백질과의 상호작용에 중요한 역할을 한다. HBA 계산과 관련하여, 각 원자의 수소 수용자가 되는 기능성을 분석하여 HBA의 수를 계산한다. 수소 수용자는 대표적으로 산소 원자(O) 또는 질소 원자(N)와 같이 수소 결합을 형성할 수 있는 원자를 의미한다. 원자의 수소 결합 능력을 파악하여 HBA의 수를 세는 방식으로 도출한다. 예를 들어, 약물로서 아세트아미노펜의 SMILES 표현은 "CC(=O)Nc1ccc(cc1)O"이다. 이를 RDKit와 같은 화학 정보 도구로 읽어온 후 수소 수용자가 되는 기능성 원자를 확인하면, 아세트아미노펜 분자 내에 수소 수용자가 되는 산소 원자가 4개라는 결과를 얻을 수 있다. 따라서 이 경우 HBA의 수는 4가 되고, 약물 데이터에서 도출된 6개 이상의 추가 입력값 중 하나로 도출할 수 있다.

[0109] HBD(수소결합 기공, Hydrogen Bond Donor)는 약물 분자가 수소 결합을 형성할 수 있는 기공의 수를 의미한다. 수소 결합은 약물 분자가 단백질과 상호작용하는데 중요한 역할을 할 수 있으며, 약물의 생체 내 활동성과 특성을 예측하는데 사용된다.

[0110] 본 실시예에 따른 청소율 예측부(150)의 HBD 추가 입력값 도출과 관련하여 부연 설명하면, 약물 데이터에서 HBD (수소결합 기공, Hydrogen Bond Donor)의 입력값을 도출하기 위해서는 해당 약물의 분자 구조를 분석하여 수소 공여자가 되는 기능성 원자의 수를 계산해야 한다. 이를 수행하기 위해 화학 정보 도구나 라이브러리를 사용할 수 있다. 아래는 예시적인 절차로 다음과 같다.

[0111] 위에서 언급한 바와 같이, 약물 데이터로 주어진 분자 구조를 SMILES 형식으로 변환하고, SMILES로 표현된 분자 구조를 화학 정보 도구나 라이브러리를 사용하여 SMILES로 표현된 분자 구조를 화학 정보 도구나 라이브러리를 사용하여 읽는다. 대표적으로 RDKit (Cheminformatics Software) 등이 사용될 수 있다. 수소 공여자 확인과 관련하여, 읽어온 분자 구조를 기반으로 각 원자들의 수소 결합을 분석한다. 수소 공여자는 다른 원자에게 수소 원자의 비공유 전자쌍을 공여하여 수소 결합을 형성하는 기능성 원자를 의미한다. HBD 계산과 관련하여, 각 원자의 수소 공여자가 되는 기능성을 분석하여 HBD의 수를 계산한다. 수소 공여자는 주로 수소 원자가 결합하는 수소결합 가능한 원자를 나타내며, 이를 세는 방식으로 HBD의 수를 도출한다. 예를 들어, 약물로서 아세트아미노펜의 SMILES 표현은 "CC(=O)Nc1ccc(cc1)O"이다. 이를 RDKit와 같은 화학 정보 도구로 읽어온 후 수소 공여자가 되는 기능성 원자를 확인하면, 아세트아미노펜 분자 내에 수소 공여자가 되는 질소 원자가 2개라는 결과를 얻을 수 있다. 따라서 이 경우 HBD의 수는 2가 될 수 있고, 약물 데이터에서 도출된 6개 이상의 추가 입력값 중 하나로 도출할 수 있다.

[0112] Refractivity(굴절률)은 분자의 굴절률을 나타내는 값으로, 약물 분자의 굴절률은 분자의 화학 구조와 물성과 관련이 있다. 이 값은 약물의 생체 내 분포와 약물-단백질 상호작용에 영향을 미칠 수 있다.

[0113] 위에서 언급한 바와 같이, 약물 데이터로 주어진 분자 구조를 SMILES 형식으로 변환하고, SMILES로 표현된 분자 구조를 화학 정보 도구나 라이브러리를 사용하여 읽어온 분자 구조를 기반으로 굴절률 값을 계산한다. 예를 들어, "CC(=O)Nc1ccc(cc1)O"라는 SMILES 문자열을 RDKit를 사용하여 분자 객체로 변환한 후, 3D 좌표를 생성하고 해당 약물의 분자 구조에 따른 굴절률 값을 계산하여 약물 데이터에서 도출된 6개 이상의 추가 입력값 중 하나로 도출할 수 있다.

[0114] PSA(극성 표면적, Polar Surface Area)은 분자의 극성 표면적을 나타내는 값으로, 극성 표면적이 높을수록 약물 분자의 수용성이 증가하며, 생체 내 분포와 대사에 영향을 미칠 수 있다. 약물 데이터로 주어진 분자 구조를 SMILES 형식으로 변환하고, SMILES로 표현된 분자 구조를 화학 정보 도구나 라이브러리를 사용하여 읽어온 분자 구조를 기반으로 분자 내 극성 원자의 표면적을 계산하고 이를 합산하여 극성 표면적 값을 약물 데이터에서 도출된 6개 이상의 추가 입력값 중 하나로 도출할 수 있다.

[0115] Log p(옥탄올/물 분배 계수, Octanol/Water Partition Coefficient): 약물 분자의 옥탄올과 물 간의 분배 상태를 나타내는 값으로, 약물의 지질용해성과 물용해성을 나타내는 중요한 특성이다. 이 값이 높을수록 지질 용해성이 높아지며, 약물의 생체 내 분포와 대사에 영향을 미칠 수 있다. 약물 데이터로 주어진 분자 구조를 SMILES 형식으로 변환하고, SMILES로 표현된 분자 구조를 화학 정보 도구나 라이브러리를 사용하여 읽어온 분자 구조를 기반으로 하여, 분자의 구조와 물성 사이의 상관 관계를 통해 Log p 값을 계산하여 약물 데이터에서 도출된 6개 이상의 추가 입력값 중 하나로 도출할 수 있다.

[0116] 본 실시예에 따른 청소율 예측부(150)는 결합확률 예측부(140)에서 추출된 결합 확률에 추가 입력값을 결합하여 통계 기반 모델에 학습하여 청소율을 예측하는 구성으로 다음의 과정을 거친다.

[0117] 먼저, 청소율 예측부(150)는 약물 시퀀스 데이터와 표적 단백질 시퀀스 데이터로부터 필요한 추가 입력값인 Rotatable Bond Count, HBA (Hydrogen Bond Acceptor), HBD (Hydrogen Bond Donor), Refractivity, PSA (Polar Surface Area), Log p (Octanol/Water Partition Coefficient) 등을 추출한다. 이들 추가 입력값은 약물의 화학적 특성을 나타내는 수치로, 약물의 특성을 더 잘 설명하기 위해 사용된다.

[0118] 다음으로 결합확률 예측부 결과 추출과 관련하여, 결합확률 예측부(140)에서 학습된 인공신경망 모델은 약물과 표적 단백질의 시퀀스 데이터를 입력으로 받아 결합 확률을 예측한다. 이 모델을 사용하여 새로운 약물과 표적 단백질 시퀀스 데이터에 대한 결합 확률을 예측한다.

[0119] 다음으로 청소율 예측부(150)는 통계 기반 모델 학습과 관련하여, 약물 데이터에서 도출된 추가 입력값들과 결합확률 예측부에서 추출된 결합 확률을 합친다. 이러한 데이터를 사용하여 통계 기반 모델을 학습한다. 통계 기반 모델은 추가 입력값과 결합 확률 사이의 관계를 학습하고, 이를 바탕으로 청소율을 예측하는데 사용된다.

[0120] 청소율 예측부(150)의 청소율 예측과 관련하여, 앞서 학습된 통계 기반 모델은 새로운 약물 데이터의 추가 입력값과 결합확률 예측부에서 추출된 결합 확률을 입력으로 받아 청소율을 예측한다. 이러한 예측된 청소율과 실제 청소율 데이터를 비교하여 모델의 성능을 평가할 수 있다. 평균 제곱 오차(Mean Squared Error, MSE) 등의 손실 함수를 사용하여 예측 값과 실제 값의 오차를 계산하고, 이를 최소화하는 방향으로 모델을 최적화한다.

[0121] 도 3은 본 발명의 일 실시예에 따른 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템에서, 약물과 단백질 시퀀스의 입력에 따라 청소율을 예측하는 방법을 순서대로 도시한 플로우 차트이다.

[0122] S110은 수집된 약물 시퀀스 데이터와 단백질 시퀀스 데이터를 서브워드 토크나이저를 사용해 단어 집합을 생성하는 토크나이징 단계에 해당된다. S111은 만들어진 단어 집합을 사용하여 학습모델에 학습시키기 위해 시퀀스 데이터를 임베딩 해주는 임베딩 단계에 해당된다. S112는 패딩 토큰을 활용하여 길이가 서로 다른 약물 시퀀스 데이터와 표적 단백질 시퀀스 데이터의 길이를 통일시켜주는 데이터 길이 통일 단계에 해당된다. S120은 전처리가 완료된 데이터를 인공신경망을 통해 학습하여 약물과 표적 단백질의 결합확률을 예측하는 결합확률 예측 단계에 해당된다. S210은 추출된 결합 확률에 약물 데이터에서 도출된 6개 이상의 추가 입력값들을 추가하여 통계 기반 모델에 학습하여 청소율을 예측하는 청소율 예측 단계에 해당된다. 이러한 순서를 통해 최종 목표인 청소율을 예측할 수 있다. 결과값 도출에 사용된 Model1의 내부 구조는 도 4와 같다.

[0123] 도 4는 본 발명의 일 실시예에 따른 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템의 결과값 도출에 사용된 Model1의 내부 구조를 나타낸 도면이다.

[0124] 도 4에 도시된 바와 같이, 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템의 결과값 도출에 사용된 모델의 내부 구조를 보면 약물 시퀀스 데이터인 SMILES 데이터를 입력으로 받는 모델과 단백질 시퀀스 데이터를 입력으로 받는 모델 2개로 나누어져 있다. 각각의 모델에 pretrained된 모델을 사용하였는데 각각 RoBERTa, ProtBert라는 모델을 사용하여 추론을 진행하였다. Model1과 Model2를 거쳐 Clearance를 추론한 결괏값은 다음 표와 같다.

[0125] 표 1

| | | Random Forest | | XGBoost | |
|---|---|---|---|---|---|
| | | Validation | Test | Validation | Test |
| Binding probability + 4 Features | $R^2$ | 0.987 | 0.990 | 0.981 | 0.989 |
| | MSE | 10.66 | 8.06 | 15.20 | 9.23 |
| | MAE | 1.355 | 0.717 | 2.221 | 1.589 |
| | RMSE | 3.260 | 2.830 | 3.899 | 3.039 |

[0126]

[0127] 검증 데이터셋에서는 MAE가 1.355의 오차로 예측하였고 테스트 데이터셋에서는 0.717의 오차로 예측함으로써 상당히 정확하게 청소율을 예측한 것으로 확인되었다.

[0128] 참고로, SMILES 데이터는 간단한 화학 소자의 정보 표현 방법을 나타내는 영어 약어로, Simplified Molecular Input Line Entry System의 약자이다. SMILES는 화학 소자(분자 또는 화합물)를 텍스트 형식으로 표현하는데 사용되며, 컴퓨터 프로그램이나 데이터베이스에서 화학 정보를 저장하고 검색하는데 주로 사용된다. SMILES 문자열은 화학 소자의 원자, 결합, 수소 원자, 분자의 방향성 등을 간단한 텍스트 형태로 표현하는 것으로, 화학적 구조를 문자열로 변환하여 저장하고 처리하는 데 편리함을 제공한다. 이러한 문자열 표현은 화학적 특성을 표현하고 화학적 구조를 기계적으로 다룰 수 있게 한다. 예를 들어, 이산화탄소(CO2)의 SMILES 표현은 "O=C=O"이다. 이러한 SMILES 표현은 화학 소자의 구조를 간단하게 표현하며, 약물 또는 화합물의 화학 구조를 기계적으로 다룰 때 유용하게 사용된다. 이와 같은 SMILES 데이터는 약물 데이터나 화학 정보를 기계 학습 모델에 입력으로 사용된다.

[0129] 본 실시예에서의 pretrained된 모델은 미리 학습된 가중치와 편향을 가진 인공신경망 모델을 말한다. 이 모델은 대규모의 데이터셋을 사용하여 사전에 학습되어 일반적인 특징과 패턴을 인식하고 내재된 정보를 포착한 상태이다. 이렇게 사전에 학습된 가중치와 편향을 가진 모델을 사용하면 새로운 태스크나 작업에 대해 더 빠르게 학습하고, 보다 효과적인 성능을 얻을 수 있다.

[0130] Pretrained 모델은 큰 규모의 데이터셋에서 대규모의 컴퓨팅 리소스를 사용하여 미리 학습된 다양한 딥러닝 아키텍처로 구성된다. 흔히 사용되는 pretrained 모델들로는 BERT(Bidirectional Encoder Representations from Transformers), GPT(Generative Pre-trained Transformer), RoBERTa(A Robustly Optimized BERT Pretraining Approach), ProtBert(Protein-BERT (A BERT-based model for protein sequences)) 등이 있으며, 이들 모델은 텍스트 처리, 자연어 이해, 이미지 분류 등 다양한 분야에서 사용된다.

[0131] 예를 들어, pretrained된 언어 모델인 BERT는 대량의 자연어 텍스트 데이터를 사용하여 학습되며, 자연어 이해, 감성 분석, 기계 번역 등의 다양한 자연어 처리 작업에 사용될 수 있다. pretrained 모델은 데이터를 다시 학습하는 것이 아니라 미리 학습된 가중치와 편향을 활용하여 새로운 작업에 적용하는 방식으로 사용된다. 이를 전이 학습(Transfer Learning)이라고도 한다.

[0132] 도 5는 본 발명의 일 실시예에 따른 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 방법을 나타낸 흐름도이다.

[0133] 도 5에 도시된 바와 같이, 인공신경망 기반의 약물-표적 단백질 상호작용 모델을 활용한 청소율 예측 시스템(이하, 청소율 예측 시스템이라 칭함)을 이용한 방법은 우선, (a)청소율 예측 시스템이 수집된 약물 시퀀스 데이터와 단백질 시퀀스 데이터를 서브워드 토크나이저를 사용해 단어 집합을 생성한다.

[0134] 이후, (b)청소율 예측 시스템은 상기 (a)단계에서 만들어진 단어 집합을 사용하여 학습모델에 학습시키기 위해 시퀀스 데이터를 임베딩한다.

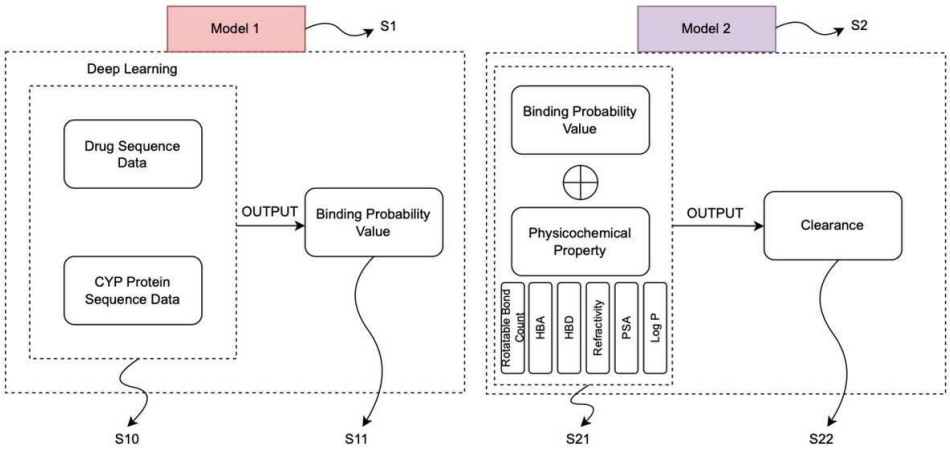
[0135] 다음으로 (c)청소율 예측 시스템은 상기 (b)단계의 임베딩 결과, 패딩 토큰을 활용하여 길이가 서로 다른 약물 시퀀스 데이터와 표적 단백질 시퀀스 데이터의 길이를 통일시켜 전처리한다.

[0136] 다음으로 (d)청소율 예측 시스템은 전처리가 완료된 데이터를 인공신경망을 통해 학습하여 약물과 표적 단백질의 결합확률을 예측한다.

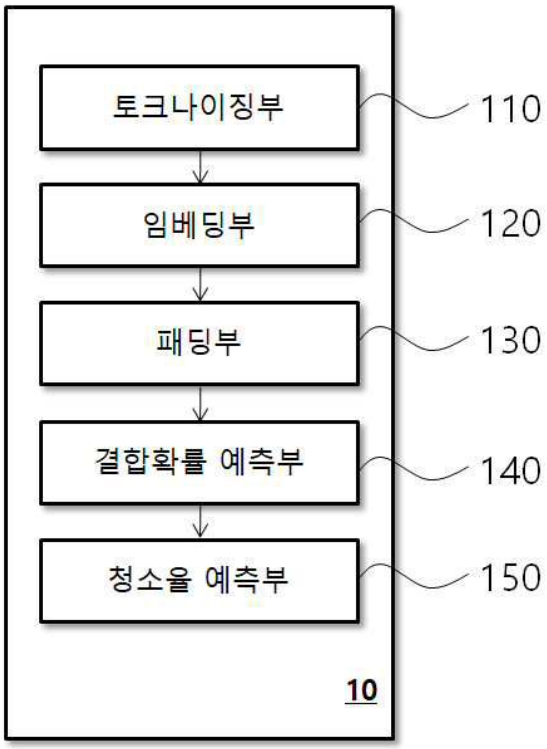
[0137] 다음으로 (e)청소율 예측 시스템은 약물데이터에서 도출된 추가 입력값들을 추가하여 투여하는 약물과 그 약물의 표적이 되는 단백질의 상호작용을 통해 구해진 결합 확률값을 가지고 인공지능을 사용하여 청소율을 예측한다.

## 도면

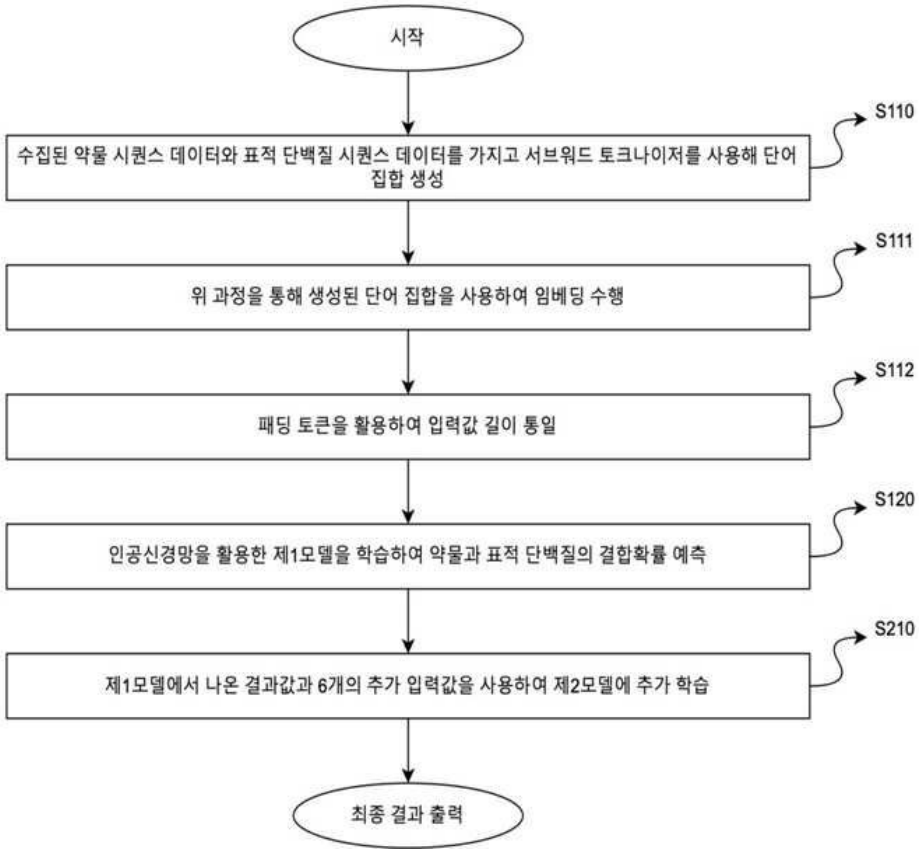
### 도면1


Model 1
S1
Deep Learning
Drug Sequence Data
CYP Protein Sequence Data
OUTPUT
Binding Probability Value
S10
S11
Model 2
S2
Binding Probability Value
Physicochemical Property
Rotatable Bond Count
HBA
HBD
Refractivity
PSA
Log P
OUTPUT
Clearance
S21
S22

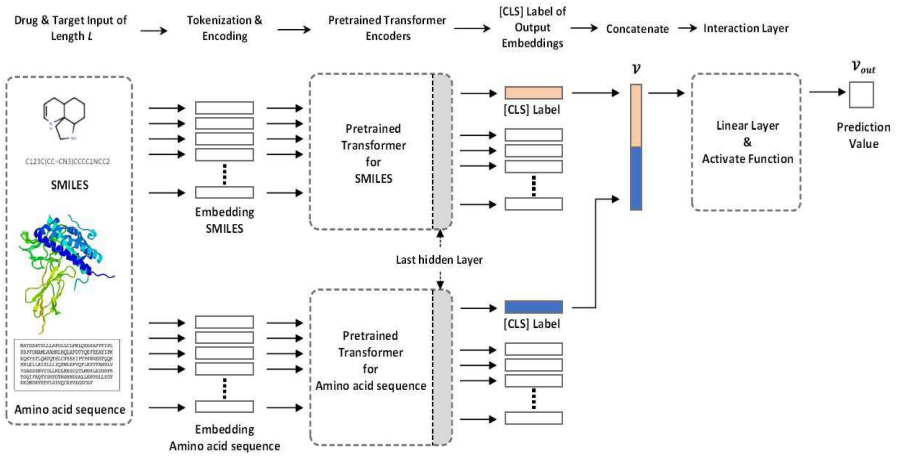
### 도면2


토크나이징부
110
임베딩부
120
패딩부
130
결합확률 예측부
140
청소율 예측부
150
10

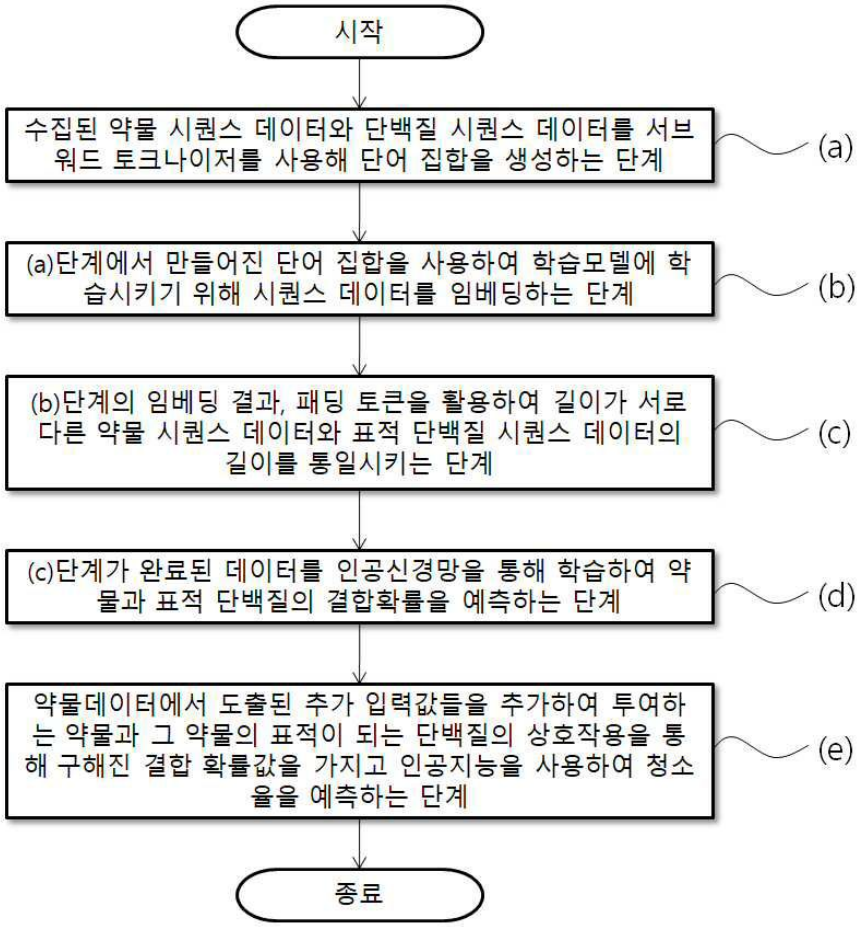
도면3


시작
수집된 약물 시퀀스 데이터와 표적 단백질 시퀀스 데이터를 가지고 서브워드 토크나이저를 사용해 단어 집합 생성
S110
위 과정을 통해 생성된 단어 집합을 사용하여 임베딩 수행
S111
패딩 토큰을 활용하여 입력값 길이 통일
S112
인공신경망을 활용한 제1모델을 학습하여 약물과 표적 단백질의 결합확률 예측
S120
제1모델에서 나온 결과값과 6개의 추가 입력값을 사용하여 제2모델에 추가 학습
S210
최종 결과 출력

도면4


Drug & Target Input of Length L
Tokenization & Encoding
Pretrained Transformer Encoders
[CLS] Label of Output Embeddings
Concatenate
Interaction Layer
SMILES
Embedding SMILES
Pretrained Transformer for SMILES
[CLS] Label
Last hidden Layer
Amino acid sequence
Embedding Amino acid sequence
Pretrained Transformer for Amino acid sequence
[CLS] Label
v
Linear Layer & Activate Function
$v_{out}$
Prediction Value

도면5


시작
수집된 약물 시퀀스 데이터와 단백질 시퀀스 데이터를 서브 워드 토크나이저를 사용해 단어 집합을 생성하는 단계
(a)
(a)단계에서 만들어진 단어 집합을 사용하여 학습모델에 학습시키기 위해 시퀀스 데이터를 임베딩하는 단계
(b)
(b)단계의 임베딩 결과, 패딩 토큰을 활용하여 길이가 서로 다른 약물 시퀀스 데이터와 표적 단백질 시퀀스 데이터의 길이를 통일시키는 단계
(c)
(c)단계가 완료된 데이터를 인공신경망을 통해 학습하여 약물과 표적 단백질의 결합확률을 예측하는 단계
(d)
약물데이터에서 도출된 추가 입력값들을 추가하여 투여하는 약물과 그 약물의 표적이 되는 단백질의 상호작용을 통해 구해진 결합 확률값을 가지고 인공지능을 사용하여 청소율을 예측하는 단계
(e)
종료