

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2021-0086495
(43) 공개일자 2021년07월08일

(51) 국제특허분류(Int. Cl.)

G16C 60/00 (2019.01) G16C 20/30 (2019.01)
G16C 20/50 (2019.01) G16H 20/10 (2018.01)
G16H 50/70 (2018.01)

(52) CPC특허분류

G16C 60/00 (2019.02)
G16C 20/30 (2019.02)

(21) 출원번호 10-2020-0177206

(22) 출원일자 2020년12월17일

심사청구일자 2020년12월17일

(30) 우선권주장

1020190179469 2019년12월31일 대한민국(KR)

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

강재우

서울특별시 중구 청계천로 400, 104동 1811호 (황학동, 롯데캐슬베네치아)

전민지

서울특별시 송파구 잠실로 62, 346동 1802호 (잠실동, 트리지움)

(뒷면에 계속)

(74) 대리인

특허법인엠에이피에스

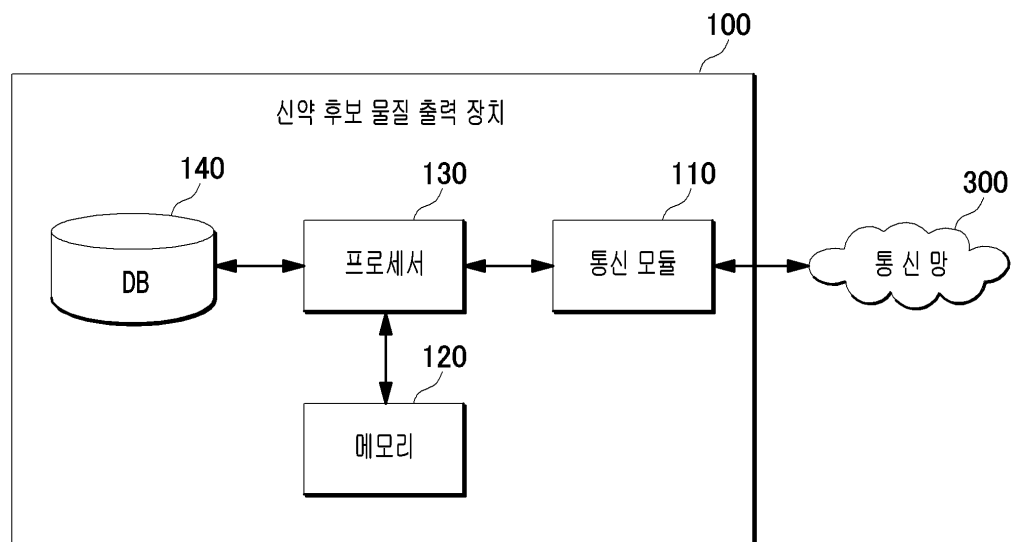
전체 청구항 수 : 총 9 항

(54) 발명의 명칭 신약 후보 물질의 출력 방법 및 장치

(57) 요약

본 발명의 일 실시예에 따른 신약 후보 물질 출력 장치는 통신 모듈; 신약 후보 물질 출력 프로그램이 저장된 메모리; 상기 신약 후보 물질 출력 프로그램을 실행하는 프로세스를 포함하되, 상기 신약 후보 물질 출력 프로그램은 화학 합성물의 화학 구조에 대한 임베딩 벡터와 각 화학 합성물에 의해 유도된 전사체량의 변화 정보에 대한 임베딩 벡터가 동일한 벡터 공간에 위치한 약물 학습 모델을 제공하고, 상기 약물 학습 모델에 입력된 신규 물질의 화학 구조에 대한 임베딩 벡터에 매칭되는 전사체량 변화 정보 결과를 출력하거나, 상기 약물 학습 모델에 입력된 타겟이 되는 전사체량 변화 정보에 매칭되는 하나 이상의 약물에 대한 정보를 출력하는 것이다.

대표도 - 도1



(52) CPC특허분류

G16C 20/50 (2019.02)

G16H 20/10 (2018.01)

G16H 50/70 (2018.01)

(72) 발명자

장부루

서울특별시 성북구 북악산로 828, 101동 505호 (정릉동, 정릉스카이쌍용아파트)

박정수

서울특별시 송파구 동남로 193, 201동 502호 (가락동, 가락쌍용아파트)

박성준

서울특별시 성북구 인촌로22길 25, 304호 (안암동 5가, 더반하우스)

김선규

서울특별시 성북구 인촌로7가길 57, 101호 (안암동 2가)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711118498
과제번호	2020R1A2C3010638
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	중견연구자지원사업 (유형2)중견연구
연구과제명	생명의료 빅데이터와 딥러닝 기반 데이터주도 신약후보물질 발굴
기 여 율	6/10
과제수행기관명	고려대학교
연구기간	2020.03.01 ~ 2021.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호	1711116320
과제번호	2017M3C4A7065887
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	차세대정보·컴퓨팅기술개발(R&D)
연구과제명	실험 및 문헌정보 거대 복잡형 데이터 통합분석추론 연구
기 여 율	2/10
과제수행기관명	서울대학교
연구기간	2020.04.01 ~ 2020.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1465031667
과제번호	HR20C0021030020
부처명	보건복지부
과제관리(전문)기관명	한국보건산업진흥원
연구사업명	연구중심병원육성(R&D)
연구과제명	인공지능·빅데이터 기반 데이터주도 의료기술 개발 및 고도화
기 여 율	2/10
과제수행기관명	고려중앙학원
연구기간	2020.07.16 ~ 2020.12.31

명세서

청구범위

청구항 1

신약 후보 물질 출력 장치에 있어서,

통신 모듈;

신약 후보 물질 출력 프로그램이 저장된 메모리;

상기 신약 후보 물질 출력 프로그램을 실행하는 프로세스를 포함하되,

상기 신약 후보 물질 출력 프로그램은 화학 합성물의 화학 구조에 대한 임베딩 벡터와 각 화학 합성물에 의해 유도된 전사체량의 변화 정보에 대한 임베딩 벡터가 동일한 벡터 공간에 위치한 약물 학습 모델을 제공하고, 상기 약물 학습 모델에 입력된 신규 물질의 화학 구조에 대한 임베딩 벡터에 매칭되는 전사체량 변화 정보 결과를 출력하거나, 상기 약물 학습 모델에 입력된 타겟이 되는 전사체량 변화 정보에 매칭되는 하나 이상의 약물에 대한 정보를 출력하는 것인, 신약 후보 물질 출력 장치.

청구항 2

제 1 항에 있어서,

상기 약물 학습 모델은 삼중항 손실 함수를 통해, 전사체량의 변화 정보를 나타내는 제 1 임베딩 벡터와 해당 전사체량 변화를 유도하는 화학 합성물의 임베딩 벡터간의 거리는 최소화하고, 상기 제 1 임베딩 벡터와 해당 전사체량의 변화를 유도하지 않는 화학 합성물의 임베딩 벡터와의 거리는 최대화하도록 구성된 것인, 신약 후보 물질 출력 장치.

청구항 3

제 1 항에 있어서,

상기 약물 학습 모델은 상기 화학 합성물의 투약전 전사체량에 대한 데이터가 입력되었을 때, 상기 약물 학습 모델이 출력하는 변화된 전사체량의 데이터와 실제 화학 합성물 투약 후 전사체량의 데이터간의 차이가 최소화 되도록 반복 학습이 진행된 것인, 신약 후보 물질 출력 장치.

청구항 4

신약 후보 물질 발굴을 위한 신약 후보 물질 출력 장치의 약물 학습 모델 구축 방법에 있어서,

화학 합성물의 화학 구조에 대한 임베딩 벡터와 각 화학 합성물에 의해 유도된 전사체량의 변화 정보에 대한 임베딩 벡터가 동일한 벡터 공간에 배치시킨 약물 학습 모델을 구축하는 단계; 및

상기 화학 합성물의 투약전 전사체량에 대한 데이터가 상기 약물 학습 모델에 입력되었을 때, 상기 약물 학습 모델이 추론하는 전사체량의 데이터와 실제 화학 합성물 투약 후 전사체량의 데이터간의 차이가 최소화되도록 반복 학습을 수행하는 단계를 포함하는, 약물 학습 모델 구축 방법.

청구항 5

제 4 항에 있어서,

상기 약물 학습 모델을 구축하는 단계는

삼중항 손실 함수를 통해, 전사체량의 변화 정보를 나타내는 제 1 임베딩 벡터와 해당 전사체량 변화를 유도하는 화학 합성물의 임베딩 벡터간의 거리는 최소화하고, 상기 제 1 임베딩 벡터와 해당 전사체량의 변화를 유도하지 않는 화학 합성물의 임베딩 벡터와의 거리는 최대화하는 것인, 약물 학습 모델 구축 방법.

청구항 6

신약 후보 물질 출력 장치를 이용한 신약 후보 물질 출력 방법에 있어서,

화학 합성물의 화학 구조에 대한 임베딩 벡터와 각 화학 합성물에 의해 유도된 전사체량의 변화 정보에 대한 임베딩 벡터가 동일한 벡터 공간에 위치한 약물 학습 모델이 제공되는 단계 및

상기 약물 학습 모델에 신규 물질의 화학 구조에 대한 임베딩 벡터 또는 타겟이 되는 전사체량 변화 정보가 입력되는 단계;

상기 약물 학습 모델이 상기 신규 물질의 화학 구조에 대한 임베딩 벡터에 매칭되는 전사체량 변화 정보 결과를 출력하거나, 상기 타겟이 되는 전사체량 변화 정보에 매칭되는 하나 이상의 약물에 대한 정보를 출력하는 단계를 포함하는, 신약 후보 물질 출력 방법.

청구항 7

제 6 항에 있어서,

상기 약물 학습 모델은 삼중항 손실 함수를 통해, 전사체량의 변화 정보를 나타내는 제 1 임베딩 벡터와 해당 전사체량 변화를 유도하는 화학 합성물의 임베딩 벡터간의 거리는 최소화하고, 상기 제 1 임베딩 벡터와 해당 전사체량의 변화를 유도하지 않는 화학 합성물의 임베딩 벡터와의 거리는 최대화하도록 구성된 것인, 신약 후보 물질 출력 방법.

청구항 8

제 6 항에 있어서,

상기 약물 학습 모델은 상기 화학 합성물의 투약전 전사체량에 대한 데이터가 입력되었을 때, 상기 약물 학습 모델이 출력하는 변화된 전사체량의 데이터와 실제 화학 합성물 투약 후 전사체량의 데이터간의 차이가 최소화 되도록 반복 학습이 진행된 것인, 신약 후보 물질 출력 방법.

청구항 9

제 4 항 내지 제 8 항 중 어느 한 항에 해당하는 방법을 수행하는 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록 매체.

발명의 설명

기술 분야

[0001] 본 발명은 전사체 표현형(transcriptome phenotype)을 이용하여 신약 후보 물질을 도출하기 위한 신약 후보 물질의 출력 방법 및 장치에 관한 것이다.

배경 기술

[0002] 신약은 신약 발견 단계와 신약 개발 단계로 구성된 프로세스에 의해 개발된다. 신약 발견 단계는 타겟 확인, 후보 물질 설계, 효능 측정, 약 후보 물질 선택을 포함한다. 신약 개발 단계에는 안전성 평가와 약물 후보자의 임상 시험이 포함된다. 신약 발견 단계와 신약 개발 단계를 통해 약품을 상용화하는 데는 상당한 시간과 비용이 소모되지만, 그 성공률은 높지 않은 것으로 알려져 있다.

[0003] 신약 개발 파이프 라인에서 질병에 적합한 타겟 단백질을 확인하고 타겟에 결합하는 분자를 찾는 것이 매우 중요하다. 일단 질병에 대한 타겟이 확인되면 타겟에 결합할 수 있는 화합물이 고효율 스크리닝을 통해 발견되고, 타겟에 결합하는 약물의 구조 유사체도 약물 후보 물질로 선택된다.

[0004] 이렇게 약물 후보 물질로 약 5000~10000 개 이상이 선정되지만, 실험과 검증을 거쳐 판매되기까지의 성공률이 0.02% 미만이기 때문에 신약 개발 비용이 과 개발 시간이 많이 소요된다.

[0005] 이와 같이, 신약 개발 과정은 시간과 비용이 많이 필요할 뿐만 아니라 어려운 과정으로서 개발되는 신약이 실제로 성공할지에 대해서도 장담할 수 없다. 게다가, 제약 업계의 연구 개발 비용은 증가하고 있으며, 새로 승인된 의약품의 수에 대한 연구 개발 비용의 비율로 계산되는 생산성은 1950 년대 이후 매년 꾸준히 감소하고 있다. 신약 개발의 성공은 신약 후보 선택에 달려 있기 때문에 신약 개발 생산성을 높이기 위해 성공확률이 높은 신약

후보를 선택하는 것이 중요하다.

[0006] 신약 발견 단계에서 신약 후보 물질 발굴을 위해 사용되었던 전통적인 방식은 타겟 중심으로 신약 후보 물질을 발굴하는 방법이다. 이는 발병 관련 주요 인자를 밝혀내는 과정인 타겟 단백질 발굴, 타겟 단백질에 물리적으로 결합하여 기능을 억제할 수 있는 화합물을 찾는 과정인 유효 물질 발굴(Hit discovery), 앞서 찾아낸 유효 물질을 구조적으로 최적화하는 선도물질 최적화 과정(Lead optimization)으로 진행된다. 타겟 단백질로부터 발굴한 약물 후보군 중에서 개발(Development) 단계를 통해 세포, 조직, 개체까지의 효과가 있는 약물을 최종적으로 고르게 된다.

[0007] 하지만 1) 타겟 단백질 가설이 필수적으로 필요하다는 점, 2) 타겟 단백질 가설을 찾았더라도 타겟 단백질에 화합물이 결합할 수 없는 구조(undruggable target)를 가진다면 약물 개발이 어려울 수 있다는 점, 3) 타겟 단백질에 화합물이 결합할 수 있는 구조라 하더라도 무수히 많은 화합물을 실험적으로 검증하기 어렵고, 후보물질 도출에 약 5.5년 이상의 상당한 시간이 소요된다는 점이 타겟 중심 신약개발 과정의 한계이다. 구체적인 예는 아래와 같다.

[0008] 예를 들어 기존 타겟 중심 신약개발 방식은 타겟 단백질을 먼저 설정하고 해당 단백질에 결합하는 화합물을 탐색한다. 하지만 해당 질병에 대한 이해도가 높지 않아 타겟 단백질을 알아낼 수 없다면, 해당 질병을 치료할 수 있는 신약개발 프로세스를 시작할 수조차 없다. 가령 알츠하이머 질병 같은 경우, 해당 질병의 요인으로 작용하는 명확한 타겟 단백질을 밝혀내지 못하여 신약후보물질을 탐색하는데 어려움이 있다.

[0009] 또한 해당 질병 치료에 중요한 역할을 하는 타겟 단백질은 알지만 실제로 그 단백질에 결합하는 물질을 만들 수 없는 경우도 약물 개발이 어려울 수 있다. 폐암이나 대장암의 타겟 단백질로 널리 알려진 KRAS는 약물이 결합할 수 있는 바인딩 사이트(binding site)가 없어 KRAS 억제제(inhibitor)가 현재 존재하지 않는다. 이는 해당 질병에 대한 이해도가 아무리 높아져도 기존 타겟 중심 방식으로는 이 질병에 대한 신약개발이 불가능함을 의미한다.

선행기술문헌

특허문헌

[0010] (특허문헌 0001) 대한민국 공개특허공보 제 10-2018-0058648 호(발명의 명칭 : 비구조-구조 전이 부위를 표적으로 하는 신약 후보 물질 발굴 방법 및 신약 후보 물질 발굴 장치)

발명의 내용

해결하려는 과제

[0011] 본 발명은 전술한 문제점을 해결하기 위하여, 본 발명의 일 실시예에 따라 전사체량 변화 데이터 및 각 화학 합성물 데이터를 입력으로 학습된 약물 학습 모델을 통해 신약 후보 물질을 출력할 수 있는 신약 후보 물질 출력 장치 및 방법을 제공하는 것에 그 목적이 있다.

[0012] 다만, 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

[0013] 상기한 기술적 과제를 달성하기 위한 기술적 수단으로서 본 발명의 일 실시예에 따른 신약 후보 물질 출력 장치는 통신 모듈; 신약 후보 물질 출력 프로그램이 저장된 메모리; 상기 신약 후보 물질 출력 프로그램을 실행하는 프로세스를 포함하되, 상기 신약 후보 물질 출력 프로그램은 화학 합성물의 화학 구조에 대한 임베딩 벡터와 각 화학 합성물에 의해 유도된 전사체량의 변화 정보에 대한 임베딩 벡터가 동일한 벡터 공간에 위치한 약물 학습 모델을 제공하고, 상기 약물 학습 모델에 입력된 신규 물질의 화학 구조에 대한 임베딩 벡터에 매칭되는 전사체량 변화 정보 결과를 출력하거나, 상기 약물 학습 모델에 입력된 타겟이 되는 전사체량 변화 정보에 매칭되는 하나 이상의 약물에 대한 정보를 출력하는 것이다.

[0014] 또한, 본 발명의 다른 실시예에 따른 신약 후보 물질 발굴을 위한 학습 모델 구축 방법은, 화학 합성물의 화학 구조에 대한 임베딩 벡터와 각 화학 합성물에 의해 유도된 전사체량의 변화 정보에 대한 임베딩 벡터가 동일한

벡터 공간에 배치시킨 약물 학습 모델을 구축하는 단계; 및 상기 화학 합성물의 투약전 전사체량에 대한 데이터가 상기 약물 학습 모델에 입력되었을 때, 상기 약물 학습 모델이 추론하는 전사체량의 데이터와 실제 화학 합성물 투약 후 전사체량의 데이터간의 차이가 최소화되도록 반복 학습을 수행하는 단계를 포함한다.

- [0015] 또한, 본 발명의 다른 실시예에 따른 신약 후보 물질 출력 방법은 화학 합성물의 화학 구조에 대한 임베딩 벡터와 각 화학 합성물에 의해 유도된 전사체량의 변화 정보에 대한 임베딩 벡터가 동일한 벡터 공간에 위치한 약물 학습 모델이 제공되는 단계 및 상기 약물 학습 모델에 신규 물질의 화학 구조에 대한 임베딩 벡터 또는 타겟이 되는 전사체량 변화 정보가 입력되는 단계; 상기 약물 학습 모델이 상기 신규 물질의 화학 구조에 대한 임베딩 벡터에 매칭되는 전사체량 변화 정보 결과를 출력하거나, 상기 타겟이 되는 전사체량 변화 정보에 매칭되는 하나 이상의 약물에 대한 정보를 출력하는 단계를 포함한다.

발명의 효과

- [0016] 진술한 본 발명의 과제 해결 수단에 의하면, 신약 후보물질을 발굴하는 디스커버리 단계에 소모되는 시간과 비용을 크게 단축시킬 수 있다.
- [0017] 또한 타겟 단백질 가설이 존재하지 않거나, 타겟 단백질은 알지만 실제 그 단백질에 결합하는 물질을 만들 수 없는 경우에도 신약개발이 가능하다.

도면의 간단한 설명

- [0018] 도 1은 본 발명의 일 실시예에 따른 신약 후보 물질 출력 장치의 구성을 도시한 도면이다.
- 도 2는 본 발명의 일 실시예에 따른 신약 후보 물질 출력 장치의 동작을 설명하기 위한 예시도이다.
- 도 3은 본 발명의 일 실시예에 따른 신약 후보 물질 출력 방법의 동작을 설명하기 위한 순서도이다.

발명을 실시하기 위한 구체적인 내용

- [0019] 아래에서는 첨부한 도면을 참조하여 본원이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본원의 실시예를 상세히 설명한다. 그러나 본원은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본원을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.
- [0020] 본원 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 "전기적으로 연결"되어 있는 경우도 포함한다.
- [0021] 본원 명세서 전체에서, 어떤 부재가 다른 부재 “상에” 위치하고 있다고 할 때, 이는 어떤 부재가 다른 부재에 접해 있는 경우뿐 아니라 두 부재 사이에 또 다른 부재가 존재하는 경우도 포함한다.
- [0022] 이하 첨부된 도면을 참고하여 본 발명의 일 실시예를 상세히 설명하기로 한다.
- [0023] 도 1은 본 발명의 일 실시예에 따른 신약 후보 물질 출력 장치의 구성을 도시한 도면이다.
- [0024] 도시된 바와 같이, 신약 후보 물질 출력 장치(100)는 통신 모듈(110), 메모리(120), 프로세서(130) 및 데이터베이스(140)를 포함할 수 있다. 신약 후보 물질 출력 장치(100)는 컴퓨팅 장치를 기본으로 구성되며, 도시되어 있지 않은, 전원부, 각종 입력 장치 및 출력 장치 등을 더 포함한다.
- [0025] 통신모듈(110)은 외부의 컴퓨팅 장치와 화학 합성물의 화학 구조에 대한 데이터나, 이와 매칭되는 것으로서 화학 합성물에 의해 유도된 전사체량의 변화 정보에 대한 데이터를 송수신할 수 있다. 통신모듈(110)은 다른 네트워크 장치와 유무선 연결을 통해 제어 신호 또는 데이터 신호와 같은 신호를 송수신하기 위해 필요한 하드웨어 및 소프트웨어를 포함하는 장치일 수 있다.
- [0026] 메모리(120)에는 신약 후보 물질 출력 프로그램이 저장된다. 신약 후보 물질 출력 프로그램은 화학 합성물의 화학 구조에 대한 임베딩 벡터와 각 화학 합성물에 의해 유도된 전사체량의 변화 정보에 대한 임베딩 벡터가 동일한 벡터 공간에 위치한 약물 학습 모델을 제공하고, 입력 장치 또는 통신모듈(110)들을 통해 입력되는 쿼리에 기반하여, 신약 후보 물질을 출력한다. 이때, 입력되는 쿼리로는 신규 물질의 화학 구조에 데이터이거나, 타겟이 되는 전사체량 변화 정보일 수 있다.
- [0027] 또한, 신약 후보 물질 출력 프로그램은 약물 학습 모델을 구축하는 로직이나, 구축된 약물 학습 모델에 대하여

새로운 학습 과정을 수행하는 학습 모델 갱신 과정등을 추가로 수행할 수 있다.

- [0028] 이러한 메모리(120)에는 신약 후보 물질 출력 장치(100)의 구동을 위한 운영 체제나 누락 데이터 예측 프로그램의 실행 과정에서 발생하는 여러 종류의 데이터가 저장된다.
- [0029] 이때, 메모리(120)는 전원이 공급되지 않아도 저장된 정보를 계속 유지하는 비휘발성 저장장치 및 저장된 정보를 유지하기 위하여 전력이 필요한 휘발성 저장장치를 통칭하는 것이다.
- [0030] 또한, 메모리(120)는 프로세서(130)가 처리하는 데이터를 일시적 또는 영구적으로 저장하는 기능을 수행할 수 있다. 여기서, 메모리(120)는 저장된 정보를 유지하기 위하여 전력이 필요한 휘발성 저장장치 외에 자기 저장 매체(magnetic storage media) 또는 플래시 저장 매체(flash storage media)를 포함할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.
- [0031] 프로세서(130)는 메모리(120)에 저장된 프로그램을 실행하되, 신약 후보 물질 출력 프로그램의 실행에 따르는 전체 과정을 제어한다. 프로세서(130)가 수행하는 각각의 동작에 대해서는 추후 보다 상세히 살펴보기로 한다.
- [0032] 이러한 프로세서(130)는 데이터를 처리할 수 있는 모든 종류의 장치를 포함할 수 있다. 예를 들어 프로그램 내에 포함된 코드 또는 명령으로 표현된 기능을 수행하기 위해 물리적으로 구조화된 회로를 갖는, 하드웨어에 내장된 데이터 처리 장치를 의미할 수 있다. 이와 같이 하드웨어에 내장된 데이터 처리 장치의 일 예로써, 마이크로프로세서(microprocessor), 중앙처리장치(central processing unit: CPU), 프로세서 코어(processor core), 멀티프로세서(multiprocessor), ASIC(application-specific integrated circuit), FPGA(field programmable gate array) 등의 처리 장치를 망라할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.
- [0033] 데이터베이스(140)는 프로세서(130)의 제어에 따라, 신약 후보 물질 출력 장치에 필요한 데이터를 저장 또는 제공한다. 이러한 데이터베이스(140)는 메모리(120)와는 별도의 구성 요소로서 포함되거나, 또는 메모리(120)의 일부 영역에 구축될 수도 있다.
- [0034] 도 2는 본 발명의 일 실시예에 따른 신약 후보 물질 출력 장치의 동작을 설명하기 위한 예시도 이고, 도 3은 본 발명의 일 실시예에 따른 신약 후보 물질 출력 방법의 동작을 설명하기 위한 순서도이다.
- [0035] 먼저, 신약 후보 물질 출력 장치(100)의 약물 학습 모델을 구축하는 방법을 살펴보기로 한다(S310).
- [0036] 신약 후보 물질 출력 장치(100)의 약물 학습 모델은, 전사체량 변화 정보를 학습하는 복수 계층의 인공 신경망과 화학 합성물의 화학 구조 정보를 학습하는 복수 계층의 인공 신경망으로 구성할 수 있다.
- [0037] 각각의 계층은 기본적인 퍼셉트론 계층(완전 연결 계층(Fully-connected layer))로 구성할 수 있으며, 화학 합성물의 화학 구조를 중심으로 임베딩 벡터를 생성하기 위해, GNN(Graph Neural Network) 등을 사용할 수 있다. 예시적으로, 전사체량 변화 정보를 학습하는 복수 계층의 인공 신경망과 화학 합성물의 화학 구조 정보를 학습하는 복수 계층의 인공 신경망 모두 4개 계층을 가지며 같은 차원의 임베딩 벡터를 출력으로 생성하는 완전 연결된 인공 신경망의 형태로 구현될 수 있다.
- [0038] 신약 후보 물질 출력 장치(100)의 약물 학습 모델은 화학 합성물의 화학 구조에 대한 임베딩 벡터와 각 화학 합성물에 의해 유도된 전사체량의 변화 정보에 대한 임베딩 벡터를 동일한 벡터 공간에 배치시킨다.
- [0039] 전사체량 변화 정보는 약물 투약 전후의 세포안의 복수의 유전자의 발현도(전사체량) 변화를 의미한다. 약물 투약 전후의 전사체량 변화에는 약물이 세포에 미치는 영향, 즉 약물과 세포안의 모든 단백질 간 및 단백질과 단백질 간의 복잡한 상호작용에 대한 정보가 포함된다. 이때, 전사체량의 변화 정도는 약물을 투여하기 전에 각 유전자의 전사체량(gene-expression)으로 구성된 가우시안 분포의 평균값으로부터 약물 투여 후 각 유전자의 전사체량까지의 거리로 정의할 수 있다. 즉, 거리값에 비례하여 변화의 정도가 증가하게 된다. 이때, 두 전사체량 간의 거리는 아래와 같은 가우시안을 통해서 구할 수 있다.

$$K(x_g, m_g) = \exp\left(-\frac{\|x_g - m_g\|^2}{2\sigma_g^2}\right)$$

[0040]

m_g

σ_g

[0041] 은 약물을 투여하기 전 유전자 g 의 전사체량 값들의 평균을 의미하고, 는 약물을 투여하기 전 유전자

$$x_g$$

g 의 전사체량 값들의 표본평균을 의미하며, x_g 는 약물을 투여한 후 유전자 g 의 전사체량 값을 의미한다.

[0042] 예를 들어, 암세포주에 대하여 약물을 투약했을때의 전사체량의 변화 정보를 담고 있는 학습 데이터를 고려할 수 있으며, 이는 978 개의 대표 유전자를 포함하고 있다고 가정한다. 이때, 특정 약물의 투약에 따른 전사체량의 변화 정보는 총 978개 유전자별로, 각 유전자의 전사체에 대한 가우시안 분포의 평균값으로부터 약물 투여 후 각 유전자의 전사체량까지의 거리값을 포함하게 된다.

[0043] 이를 이용하면 기존의 타겟 중심 신약개발 방식으로는 신약개발이 불가능했던 문제를 해소할 수 있다. 예를 들어, 질병 관련 타겟 단백질 가설이 존재하지 않은 경우, 환자군의 유전자 발현 패턴을 정상군의 유전자 발현 패턴으로 유도할 수 있는 약물 후보 물질을 발굴할 수 있다. 또한, 질병 관련 타겟 단백질 가설이 알려져 있더라도 타겟 단백질이 결합 불가능하여 약물 개발이 어려웠던 경우, 타겟 유전자 녹다운(Knockdown)으로 인한 유전자 발현도 변화를 유도할 수 있는 후보 물질을 발굴할 수 있게 된다.

[0044] 도시된 바와 같이, 화학 합성물의 화학 구조에 대한 임베딩 벡터와 해당 화학 합성물이 약물로서 주입되었을 때 발생하는 전사체량의 변화 정보에 대한 임베딩 벡터를 동일한 벡터 공간(예를 들면, 단위 초구(unit hypersphere)의 표면)에 배치하여, 약물 학습 모델(210)이 구축되도록 한다.

[0045] 이때, 약물 투약 전후의 전사체량 변화 정보에는 오프 타겟(off-target) 영향을 포함한 약물의 반응에 대한 모든 정보가 포함되어 있으므로, 전사체량 변화 정보와 그 변화를 유도하는 약물을 동일한 벡터공간의 같은 곳에 위치하도록 임베딩함으로써 약물 효과에 대한 추상적 특성을 포착할 수 있게 된다. 또한, 약물 효과에 기반한 임베딩 모듈이기 때문에, 새로운 화학구조를 갖는 약물일지라도 약물 특성에 기반하여 신약 후보 물질로 발굴될 수 있다. 또한, 본 발명의 임베딩 모듈은 약물의 화학적 구조를 입력으로 사용하기 때문에, 기존에 구조가 알려진 화학 합성물뿐만 아니라 미래에 알려질 모든 합성 가능한 화합물들에 대해서도 별도의 과정과 비용 발생 없이 약물 효과를 기반으로 하는 임베딩으로 표현할 수 있다.

[0046] 또한, 본 발명의 약물 학습 모델은 화학 합성물의 화학 구조에 대한 임베딩 벡터와 해당 화학 합성물이 약물로서 주입되었을 때 발생하는 전사체량의 변화 정보에 대한 임베딩 벡터를 동일한 벡터 공간에 구축하고, 이에 대한 손실 함수로서, 삼중항 손실(triplet loss) 함수를 사용한다. 본 손실함수에서는 특정 화학 합성물의 투약에 따라 발생하는 전사체량의 변화 정보를 나타내는 임베딩 벡터(f^a), 해당 전사체량 변화를 유도하는 화학 합성물을 나타내는 임베딩 벡터(f^p), 해당 전사체량 변화를 유도하지 않는 화학 합성물을 나타내는 임베딩 벡터(f^n)를 이용한다.

[0047] $\text{Triplet_loss(Anchor, Positive, Negative)}$

$$\text{Loss} = \sum_{i=1}^N \max(\text{Dist}(f_i^a, f_i^p) - \text{Dist}(f_i^a, f_i^n) + \alpha, 0)$$

[0048]

[0049] 즉, 삼중항 손실함수는 전사체량의 변화 정보를 나타내는 임베딩 벡터(f^a)와 해당 전사체량 변화를 유도하는 화학 합성물을 나타내는 임베딩 벡터(f^p)간의 거리에서, 전사체량의 변화 정보를 나타내는 임베딩 벡터(f^a)와 해당 전사체량 변화를 유도하지 않는 화학 합성물을 나타내는 임베딩 벡터(f^n)간의 거리를 차감한값에 기초하여 산출된다.

[0050] 이때, α 는 모델 설계자에 의해 설정되는 마진 값을 나타내며, i 는 학습 횟수 또는 샘플의 식별 번호를 나타내는 것이고, N 은 각 전사체량 변화를 유도하는 각 화학 합성물과 전사체량 변화 정보의 쌍의 개수를 나타낸다. 예를 들어, 학습 데이터가 총 21,220개의 약물 및 82개의 셀라인에 대하여 실험된 310,114 개의 전사체량 변화 정보를 포함하고 있다면, N 은 전체 전사체량 변화 정보의 개수에 해당하는 310, 114가 될 수 있다.

$$\text{Dist}(A, B)$$

[0051] 임의의 두 임베딩 벡터 A, B 에 대한 거리 $\text{Dist}(A, B)$ 는 음의 코사인 유사도(Negative cosine similarity)를 사용하며, 다음과 같이 정의된다.

$$Dist(A, B) = 1 - \frac{A \cdot B}{\|A\| \times \|B\|}$$

[0052]

[0053] 삼중항 손실 함수는, 전사체량의 변화 정보를 나타내는 임베딩 벡터와 해당 전사체량 변화를 유도하는 화학 합성물의 임베딩 벡터간의 거리는 최소화하고, 해당 전사체량의 변화를 유도하지 않는 화학 합성물의 임베딩 벡터와의 거리는 최대화하도록 구성되어 있다.

[0054] 이러한 본 발명의 손실함수의 성능을 더욱 극대화하기 위하여, 아래와 같이, 더블 삼중항 손실(Double triplet loss) 함수를 사용할 수 있다.

$$Loss = \sum_{i=1}^N \max(Dist(f_i^a, f_i^p) - Dist(f_i^a, f_i^n) + \alpha, 0) \\ + \max(Dist(f_i^p, f_i^a) - Dist(f_i^p, f_i^n) + \alpha, 0)$$

[0055]

[0056] 앞서 설명한 삼중항 손실 함수와 대비하여, 후방에 전사체량 변화를 유도하는 화학 합성물을 나타내는 임베딩 벡터(f^p)와 전사체량의 변화 정보를 나타내는 임베딩 벡터(f^a)간의 거리에서, 전사체량 변화를 유도하는 화학 합성물을 나타내는 임베딩 벡터(f^p)와 전사체량 변화를 유도하지 않는 화학 합성물을 나타내는 임베딩 벡터(f^n)간의 거리를 차감한 항이 더 추가된다.

[0057] 앞서 설명한 삼중항 손실 함수를 사용하면, 전사체량의 변화 정보를 나타내는 임베딩 벡터(f^a)와 해당 전사체량 변화를 유도하지 않는 화학 합성물을 나타내는 임베딩 벡터(f^n)간의 거리가 매우 멀 경우, 손실함수 값이 0으로 계산된다. 이러한 경우, 전사체량의 변화 정보를 나타내는 임베딩 벡터(f^a)와 해당 전사체량 변화를 유도하는 화학 합성물을 나타내는 임베딩 벡터(f^p) 간의 거리가 좁혀지지 않는 경우가 발생할 수 있다.

[0058] 더블 삼중항 손실 함수를 사용하면, 전방에 위치한 항이 0이 되더라도, 후방에 추가된 항을 통해 전사체량 변화를 유도하는 화학 합성물을 나타내는 임베딩 벡터(f^p)와 전사체량의 변화 정보를 나타내는 임베딩 벡터(f^a)간의 거리를 좁히는 것이 가능하다.

[0059] 한편, 이와 같이 구축된 학습 모델을 더욱 고도화 하기 위해, 반복 학습 단계가 추가적으로 수행될 수 있다 (220).

[0060] 즉, 화학 합성물의 투약전 전사체량에 대한 데이터가 약물 학습 모델에 입력되었을 때, 약물 학습 모델이 추론하는 전사체량의 데이터와 실제 화학 합성물 투약 후 전사체량의 데이터간의 차이가 최소화되도록 반복 학습을 수행한다. 이를 위해, 소정의 손실 함수를 사용하여, 학습 모델에서 출력한 전사체량 데이터와 약물 투약후 전사체량 데이터 간의 차이가 최소가 되도록, 학습 모델의 가중치를 갱신하는 동작을 수행할 수 있다.

[0061] 다음으로, 이와 같이 구축된 학습 모델에 대하여 신규 물질의 화학 구조에 대한 임베딩 벡터 또는 타겟이 되는 전사체량 변화 정보를 쿼리로서 입력한다(S320).

[0062] 본 발명에서는, 기존에 알려진 모든 화합물들에 대해 임베딩 벡터를 구하고 이들을 빠르게 검색할 수 있도록 색인을 미리 마련하여 DB 등에 저장한다. 기존에 알려진 화합물은 대략 Zinc15 DB에 등록되어 있는 것들 기준으로는 약 13억개 수준이다. 이 화합물들을 포함해 미래에 추가될 화합물들에 대해서도 추가로 색인 가능하도록 구현한다. 이와 같이 구축한 유사 벡터 검색 시스템을 사용하면 밀집 벡터로 표현된 여러 화합물 중에서 쿼리 벡터와 가장 유사한 화합물을 찾아낼 수 있다. 이때 쿼리 벡터는 신규 물질의 화학 구조에 대한 임베딩 벡터가 되거나, 약물 투약 후 전사체량 변화 정보에 대한 임베딩 벡터가 될 수 있다.

[0063] 예를 들어, 신약 재창출(drug repositioning)과 같이 개발 중이거나 판매가 진행 중인 약물의 새로운 의학적 용도를 개발하기 위하여 해당 약물의 임베딩 벡터를 쿼리 벡터로써 사용할 수 있고, 약물 스크리닝과 같이 요구되는 기능을 수행하는 신약 후보 물질을 뽑기 위해서 전사체량 변화 정보를 쿼리로 사용할 수 있다. 이때 대략 13억 개의 화합물에 대한 유사성 검색 소요시간은 수 초 단위로 빠른 시간 내에 약물 후보군을 선정할 수 있다.

[0064] 다음으로, 약물 학습 모델은 쿼리 벡터에 대한 출력으로서, 신규 물질의 화학 구조에 대한 임베딩 벡터에 매칭

되는 전사체량 변화 정보 결과를 출력하거나, 타겟이 되는 전사체량 변화 정보에 매칭되는 하나 이상의 약물에 대한 정보를 출력한다(S330).

[0065] 도 2에 도시된 바와 같이, 약물 학습 모델은 전사체량 변화 정보가 쿼리로서 입력되면, 이에 매칭되는 약물 후보 물질을 출력한다. 또한, 약물 학습 모델은 신규 물질의 화학 구조에 대한 임베딩 벡터가 입력되면, 이에 매칭되는 전사체량 변화 정보 결과를 출력한다.

[0066] 한편, 본 발명에서는 앞서 살펴본 바와 같이, 전사체량의 변화 정보를 기초로, 이에 매칭되는 약물 후보 물질을 출력할 수 있다. 특히, 타겟 단백질이 결정된 경우에는, 타겟 단백질의 유전자를 KO(knock out)/KD(knock down) 시키기 이전 전사체량(기준 전사체량)과 타겟 단백질의 유전자를 KO/KD 시킨 이후의 전사체량(유도 전사체량)의 차이를 기초로 전사체량 변화 정보를 특정할 수 있다. 그리고, 이렇게 얻어진 전사체량 변화 정보를 약물 학습 모델에 입력하여, 약물 후보 물질을 출력할 수 있다.

[0067] 다른 방법으로, 정상군과 타겟 질병군의 전사체량 정보가 주어진 경우에는 질병군의 전사체량(기준 전사체량)과 정상군의 전사체량(유도 전사체량)의 차이를 기초로 전사체량 변화 정보를 특정할 수 있다. 그리고, 이렇게 얻어진 전사체량 변화 정보를 약물 학습 모델에 입력하여, 약물 후보 물질을 출력할 수 있다.

[0068] 한편, 전사체량 변화량의 절대적인 값의 편차(variance)는, 값의 관측 방법에 따라 모델 학습시 사용한 전사체 변화량 값의 편차와 크게 차이가 날 수 있다. 이러한 값을 약물 학습 모델의 입력으로 그대로 사용할 경우, 모델 학습 환경과 크게 다른 데이터이기 때문에 기대했던 결과를 얻기 어려울 수 있다. 이를 해결하기 위해, 쿼리로서 입력되는 전사체량 변화 정보의 편차를 약물 학습 모델의 학습 과정에 사용된 전사체량 변화 정보 학습 데이터의 편차와 대체적으로 일치시킬 필요가 있다.

[0069] 이를 위해, 새롭게 입력되는 전사체량 변화 정보를 구성하는 기준 전사체량과 유도 전사체량에 대해서 T-테스트를 통해 T-통계량(T-statistic)의 크기 순으로 각각의 유전자들을 나열한다. 이는 새롭게 입력되는 전사체 정보에 대해서 전사체 변화량의 크기 순으로 유전자들을 나열한 것과 같다. 다음으로, 전사체 변화량의 크기 순으로 나열된 유전자들에 대해서, 학습 데이터의 전사체 변화량의 값들을 큰 순으로 매핑한다. 예를 들어, 기준 전사체량 대비 유도 전사체량에서 전사체량이 가장 많이 감소된 유전자에 대해서는, 학습 데이터의 전사체 변화량 중에서도 음의 절대값이 가장 큰 값을 부여하는 식이다. 이러한 방법을 통해 새롭게 들어온 전사체 정보에 대해서, 전사체 변화량 기준 유전자들의 순서를 유지하면서 동시에 학습 데이터의 값의 편차와 동일한 수준의 편차를 갖는 입력 데이터를 생성할 수 있다.

[0070] 약물 학습 모델은, 조정된 전사체량 변화정보의 임베딩 벡터가 입력되면, 앞서 구축한 화합물 임베딩 데이터베이스로부터 가장 가까운 벡터값을 갖는 화합물들을 검색한다. 이때, 벡터 간의 거리는 유클리드 거리(Euclidean distance) 혹은 코사인 유사도(Cosine similarity)와 같은 일반적인 거리 함수를 사용할 수 있다. 이를 기반으로, 가장 가까운 벡터값을 갖는 화합물들을 기대하는 전사체량 변화 효과를 유도하는 약물 후보로서 도출한다.

[0071] 본 발명의 일 실시예는 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터에 의해 실행가능한 명령어를 포함하는 기록 매체의 형태로도 구현될 수 있다. 컴퓨터 판독 가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발성 매체, 분리형 및 비분리형 매체를 모두 포함한다. 또한, 컴퓨터 판독 가능 매체는 컴퓨터 저장 매체를 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함한다.

[0072] 본 발명의 방법 및 시스템은 특정 실시예와 관련하여 설명되었지만, 그것들의 구성 요소 또는 동작의 일부 또는 전부는 범용 하드웨어 아키텍처를 갖는 컴퓨터 시스템을 사용하여 구현될 수 있다.

[0073] 전술한 본원의 설명은 예시를 위한 것이며, 본원이 속하는 기술분야의 통상의 지식을 가진 자는 본원의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.

[0074] 본원의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본원의 범위에 포함되는 것으로 해

석되어야 한다.

부호의 설명

[0075]

100: 신약 후보 물질 출력 장치

110: 통신 모듈

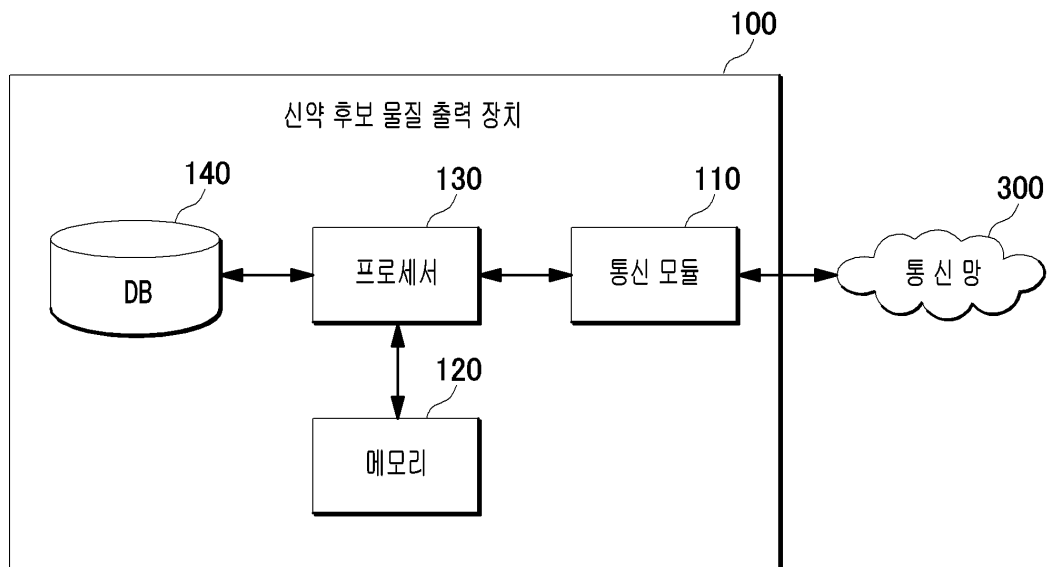
120: 메모리

130: 프로세서

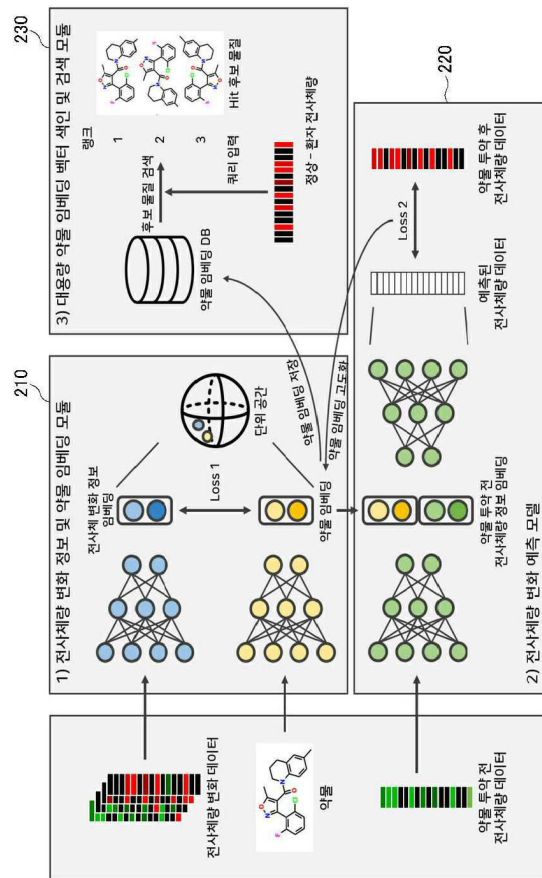
140: 데이터베이스

도면

도면1



도면2



도면3

