



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0141290
(43) 공개일자 2020년12월18일

(51) 국제특허분류(Int. Cl.)
G16B 25/10 (2019.01) C12Q 1/6809 (2018.01)
G06N 20/00 (2019.01)
(52) CPC특허분류
G16B 25/10 (2019.02)
C12Q 1/6809 (2018.05)
(21) 출원번호 10-2019-0068118
(22) 출원일자 2019년06월10일
심사청구일자 2019년06월10일

(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)
(74) 대리인
특허법인엠에이피에스

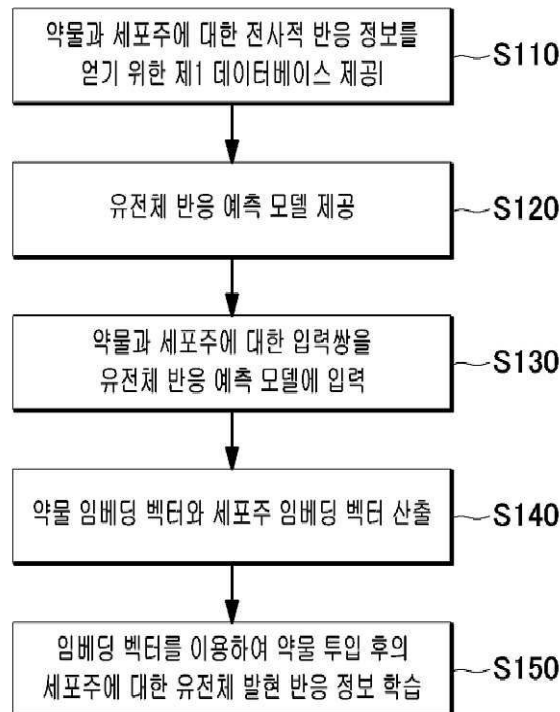
전체 청구항 수 : 총 23 항

(54) 발명의 명칭 신약 후보 물질의 발굴 시스템 및 그 방법

(57) 요약

본 발명은 신약 후보 물질의 발굴 시스템 및 그 방법에 관한 것으로서, 특히 신약 후보 물질의 발굴 방법은 기계 학습에 기반하여 신약 후보 물질을 발굴하기 위한 신약 후보 물질의 발굴 시스템에 의해 수행되는 신약 후보 물질의 발굴 방법에 있어서, a) 복수 개의 약물 또는 세포주(Cell Line)에 대한 유전자 발현 정보, 유전자 발현 역 (뒷면에 계속)

대표도 - 도2



제 정보, 각 약물의 구조 정보, 서로 다른 약물들 간의 약물 반응 유사성 정보를 포함하는 데이터베이스를 제공하는 단계; 및 b) 상기 데이터베이스에 기초하여 약물과 세포주로 구성된 입력쌍이 입력되면, 상기 입력쌍에 대해 상기 약물의 투입으로 인한 상기 세포주의 유전체 발현 반응 정보를 학습하는 유전체 반응 예측 모델을 제공하는 단계를 포함하되, 상기 유전체 반응 예측 모델은, 상기 약물의 구조 정보에 대한 약물 임베딩 벡터, 상기 세포주의 유전체 정보에 대한 세포주 임베딩 벡터를 각각 산출한 후 상기 약물 임베딩 벡터와 세포주 임베딩 벡터를 이용하여 상기 약물과 세포주에 대해 후처리된 유전체 발현 반응 정보를 학습하는 것이다.

(52) CPC특허분류

G06N 20/00 (2019.01)

(72) 발명자

이준현

서울특별시 성북구 개운사길 86-10 (안암동5가)

장부루

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

김범수

서울특별시 노원구 화랑로51길 17, 703동 1004호
(공릉동, 화랑타운아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711065336
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	(원천)바이오의료기술개발사업 (시스템·인포메틱스사업)
연구과제명	우울증 진단치료 의사결정 보조 시스템을 위한 분석 알고리즘 개발
기 여 율	1/1
과제수행기관명	고려대학교산학협력단
연구기간	2018.02.01 ~ 2018.12.31

명세서

청구범위

청구항 1

기계 학습에 기반하여 신약 후보 물질을 발굴하기 위한 신약 후보 물질의 발굴 시스템에 의해 수행되는 신약 후보 물질의 발굴 방법에 있어서,

- a) 복수 개의 약물 또는 세포주(Cell Line)에 대한 유전자 발현 정보, 유전자 발현 억제 정보, 각 약물의 구조 정보, 서로 다른 약물들 간의 약물 반응 유사성 정보를 포함하는 데이터베이스를 제공하는 단계; 및
- b) 상기 데이터베이스에 기초하여 약물과 세포주로 구성된 입력쌍이 입력되면, 상기 입력쌍에 대해 상기 약물의 투입으로 인한 상기 세포주의 유전체 발현 반응 정보를 학습하는 유전체 반응 예측 모델을 제공하는 단계를 포함하되,

상기 유전체 반응 예측 모델은, 상기 약물의 구조 정보에 대한 약물 임베딩 벡터, 상기 세포주의 유전체 정보에 대한 세포주 임베딩 벡터를 각각 산출한 후 상기 약물 임베딩 벡터와 세포주 임베딩 벡터를 이용하여 상기 약물과 세포주에 대해 후처리된 유전체 발현 반응 정보를 학습하는 것인, 신약 후보 물질의 발굴 방법.

청구항 2

제 1 항에 있어서,

상기 세포주 임베딩 벡터는 상기 세포주의 전처리된 유전자 발현 정보와 기 학습된 유전체 반응 예측 모델을 이용해 상기 약물 반응과 연관된 벡터로 임베딩되고,

상기 약물 임베딩 벡터는 상기 약물의 화학 구조의 특징을 나타내는 구조 기반 벡터와 기 학습된 유전체 반응 예측 모델을 이용해 상기 세포주의 유전체 발현 반응과 연관된 벡터로 임베딩되는, 신약 후보 물질의 발굴 방법.

청구항 3

제 1 항에 있어서,

상기 데이터베이스는, 상기 약물의 유전자 발현 정보, 상기 유전자 발현 정보로부터 예측된 차등 발현 유전자(differentially expressed gene, DEG)의 발현량을 포함한 약물 유도 전사체, 서로 다른 약물들 간의 약물 반응 유사성 정보인 CMap 점수를 포함하는 연결 지도(Connectivity Map, CMap) 데이터셋을 포함하는, 신약 후보 물질의 발굴 방법.

청구항 4

제 1 항에 있어서,

상기 유전체 반응 예측 모델은,

상기 입력쌍이 입력되면 상기 약물의 화학 구조의 특징을 나타내는 구조 기반 벡터와 상기 세포주에 대한 유전체 발현 정보, 약물 투입 정보가 입력되는 입력층;

상기 약물의 구조 기반 벡터에 기초하여 약물 임베딩 벡터를 산출하는 제1 네트워크와 상기 세포주의 유전체 발현 정보에 기초하여 세포주 임베딩 벡터를 산출하는 제2 네트워크를 포함하는 임베딩층; 및

상기 약물 투입 정보, 상기 약물 임베딩 벡터, 상기 세포주 임베딩 벡터를 결합하여 상기 약물 투입 후의 유전체 발현 반응 정보를 출력하는 예측층을 포함하는 신약 후보 물질의 발굴 방법.

청구항 5

제 4 항에 있어서,

상기 약물 투입 정보는 상기 약물의 복용 정보 또는 상기 약물 투여 기간 정보를 포함하는 것인, 신약 후보 물

질의 발굴 방법.

청구항 6

제 4 항에 있어서,

상기 유전체 반응 예측 모델은,

상기 약물 임베딩 벡터와 세포주 임베딩 벡터의 유사도를 계산하고, 상기 계산된 유사도를 이용해 상기 예측층에서 예측된 유전체 발현 반응 정보가 상기 입력층의 약물 반응 유사성 정보가 되도록 학습하는 것인, 신약 후보 물질의 발굴 방법.

청구항 7

제 4 항에 있어서,

상기 약물 임베딩 벡터(V_D)는 하기 수학식 1에 의해 생성되고, 상기 세포주 임베딩 벡터(V_C)는 하기 수학식 2에 의해 생성되는 것인, 신약 후보 물질의 발굴 방법.

[수학식 1]

$$V_D = \overline{W}_L^T (\sigma(\cdots \sigma(\overline{W}_1^T \cdot D + \overline{b}_1) \cdots) + \overline{b}_{L-1}) + \overline{b}_L,$$

[수학식 2]

$$V_C = W_L'^T (\sigma(\cdots \sigma(W_1'^T \cdot C + b'_1) \cdots) + b'_{L-1}) + b'_L,$$

V_D : 유전자의 d 차원 전사 반응 기반의 약물 임베딩 벡터, $V_D \in \mathbb{R}^d$

V_C : 유전자의 d' 차원 세포주 임베딩 벡터, $V_C \in \mathbb{R}^{d'}$

L: 은닉층의 수

$\overline{W}_i$

: 제1 네트워크의 i번째 계층의 활성화 함수

$\overline{b}_i$

: 제1 네트워크의 가중치 행렬

W_i' : 제2 네트워크의 i번째 계층의 활성화 함수

b_i' : 제2 네트워크의 가중치 행렬

σ : 바이어스

청구항 8

제 4 항에 있어서,

상기 약물 임베딩 벡터(V_D)는 SMILES(Simplified Molecular Input Line Entry System), WLN(Wiswesser line notation), InChI (International Chemical Identifier)를 포함한 화합물의 문자열 표기 방식 중 어느 하나의 방식을 이용하여 각 약물에 대한 화합물의 구조 정보를 문자열로 표기하는 것인, 신약 후보 물질의 발굴 방법.

청구항 9

제 4 항에 있어서,

상기 약물 임베딩 벡터(V_D)는 그래프 신경망(Graph Neural Networks, GNNs) 또는 접합 트리(Junction Tree) 알고리즘 중 어느 하나를 이용하여 각 약물에 대한 화합물의 구조 정보를 임베딩하는 것인, 신약 후보 물질의 발

굴 방법.

청구항 10

제 4 항에 있어서,

상기 유전체 반응 예측 모델은 다차원의 다층 신경망(Multi-Layer Perceptron, MLP)으로 구성되고,

상기 입력층은 하기 수학적 식 3에 의한 입력 데이터를 상기 임베딩층으로 입력하고, 상기 예측층은 하기 수학적 식 4

$$\hat{Y}$$

에 의해 예측된 유전체 발현 반응 정보를 출력 데이터()로 출력하는 것인, 신약 후보 물질의 발굴 방법.

[수학적 식 3]

$$X = [V_C; V_D; M; \Delta T].$$

[수학적 식 4]

$$\hat{Y} = f(C, D, M, \Delta T | \theta).$$

$$\hat{Y} = W_L^T (\sigma(\dots \sigma(W_1^T \cdot D + b_1) \dots) + b_{L-1}) + b_L,$$

V_D : 유전자의 d 차원 전사 반응 기반의 약물 임베딩 벡터, $V_D \in \mathbb{R}^d$

V_C : 유전자의 d' 차원 세포주 임베딩 벡터, $V_C \in \mathbb{R}^{d'}$

M: 약물 복용 정보

ΔT : 약물 투여 기간 정보

L: 은닉층의 수

W'_i : i 번째 계층의 활성화 함수

b'_i : 가중치 행렬

σ : 바이어스

청구항 11

제 10 항에 있어서,

상기 유전체 반응 예측 모델은 비용 함수를 하기 수학적 식 5에 의한 평균 제곱 오차(Mean square error, MSE)를 이용해 계산하고, 상기 비용 함수의 결과값이 최소화되도록 학습을 통해 최적의 가중치를 결정하는 것인, 신약 후보 물질의 발굴 방법.

[수학적 식 5]

$$MSE = \frac{1}{N} \sum_{i=1}^N \|Y_i - \hat{Y}_i\|_2^2,$$

Y_i : i 번째 입력 데이터에 대한 실제 유전체 발현 반응 값

$$\hat{Y}_i$$

: i 번째 입력 데이터에 대한 예측된 유전체 발현 반응 값

N: 데이터베이스 내의 입력 데이터의 수

청구항 12

기계 학습에 기반하여 신약 후보 물질을 발굴하기 위한 신약 후보 물질의 발굴 시스템에 의해 수행되는 신약 후보 물질의 발굴 방법에 있어서,

- a) 복수 개의 약물 또는 세포주(Cell Line)에 대한 유전자 발현 정보, 유전자 발현 억제 정보, 각 약물의 구조 정보, 서로 다른 약물들 간의 약물 반응 유사성 정보를 포함하는 제1 데이터베이스를 제공하는 단계;
- b) 상기 제1 데이터베이스에 기초하여 약물과 세포주로 구성된 입력쌍이 입력되면, 유전체 반응 예측 모델을 통해 상기 입력쌍에 대한 상기 약물의 투입으로 인한 상기 세포주의 유전체 발현 반응 정보를 학습하는 단계;
- c) 상기 유전체 발현 정보 또는 약물 반응 정보에 대해 알려지지 않은 약물 또는 세포주에 대한 특징 정보를 수집 및 관리하는 제2 데이터베이스를 제공하는 단계;
- d) 상기 제2 데이터베이스에 기초한 약물 또는 세포주가 적어도 하나 이상이 포함되도록 입력쌍을 구성하고, 상기 입력쌍을 상기 유전체 반응 예측 모델에 입력하여 상기 약물의 투입으로 인한 세포주의 유전체 발현 반응 정보를 예측하는 단계; 및
- e) 상기 예측된 유전체 발현 반응 정보가 사용자에게 의해 설정된 유전체 발현 패턴에 해당되는 경우, 상기 제2 데이터베이스에 기초한 약물 또는 세포주를 상기 사용자가 원하는 약물 활동을 유도하기 위한 후보 물질로 선정하는 단계를 포함하되,

상기 유전체 반응 예측 모델은, 상기 약물의 구조 정보에 대한 약물 임베딩 벡터, 상기 세포주의 유전체 정보에 대한 세포주 임베딩 벡터를 각각 산출한 후 상기 약물 임베딩 벡터와 세포주 임베딩 벡터를 이용하여 상기 약물과 세포주에 대해 후처리된 유전체 발현 반응 정보를 학습 및 예측하는 것인, 신약 후보 물질의 발굴 방법.

청구항 13

제 12 항에 있어서,

상기 유전체 반응 예측 모델은,

상기 입력쌍이 입력되면 상기 약물의 화학 구조의 특징을 나타내는 구조 기반 벡터와 상기 세포주에 대한 유전체 발현 정보, 약물 투입 정보가 입력되는 입력층;

상기 약물의 구조 기반 벡터에 기초하여 약물 임베딩 벡터를 산출하는 제1 네트워크와 상기 세포주의 유전체 발현 정보에 기초하여 세포주 임베딩 벡터를 산출하는 제2 네트워크를 포함하는 임베딩층; 및

상기 약물 투입 정보, 상기 약물 임베딩 벡터, 상기 세포주 임베딩 벡터를 결합하여 상기 약물 투입 후의 유전체 발현 반응 정보를 출력하는 예측층을 포함하는 신약 후보 물질의 발굴 방법.

청구항 14

제 13 항에 있어서,

상기 약물 투입 정보는 상기 약물의 복용 정보 또는 상기 약물 투여 기간 정보를 포함하는 것인, 신약 후보 물질의 발굴 방법.

청구항 15

제 13 항에 있어서,

상기 약물 임베딩 벡터(V_D)는 하기 수학적 식 1에 의해 생성되고, 상기 세포주 임베딩 벡터(V_C)는 하기 수학적 식 2에 의해 생성되는 것인, 신약 후보 물질의 발굴 방법.

[수학적 식 1]

$$V_D = \overline{W}_L^T (\sigma(\cdots \sigma(\overline{W}_1^T \cdot D + \overline{b}_1) \cdots) + \overline{b}_{L-1}) + \overline{b}_L,$$

[수학적 식 2]

$$V_C = W_L'^T (\sigma(\cdots \sigma(W_1'^T \cdot C + b'_1) \cdots) + b'_{L-1}) + b'_L,$$

V_D : 유전자의 d 차원 전사 반응 기반의 약물 임베딩 벡터, $V_D \in \mathbb{R}^d$

V_C : 유전자의 d'차원 세포주 임베딩 벡터, $V_C \in \mathbb{R}^{d'}$

L: 은닉층의 수

$\overline{w'_i}$

: 제1 네트워크의 i번째 계층의 활성화 함수

$\overline{b'_i}$

: 제1 네트워크의 가중치 행렬

w'_i : 제2 네트워크의 i번째 계층의 활성화 함수

b'_i : 제2 네트워크의 가중치 행렬

σ : 바이어스

청구항 16

제 13 항에 있어서,

상기 유전체 반응 예측 모델은 다차원의 다층 신경망(Multi-Layer Perceptron, MLP)으로 구성되고,

상기 입력층은 하기 수학적식 3에 의한 입력 데이터를 상기 임베딩층으로 입력하고, 상기 예측층은 하기 수학적식 4

$\hat{Y}$

에 의해 예측된 유전체 발현 반응 정보를 출력 데이터()로 출력하는 것인, 신약 후보 물질의 발굴 방법.

[수학적식 3]

$$X = [V_C; V_D; M; \Delta T].$$

[수학적식 4]

$$\hat{Y} = f(C, D, M, \Delta T | \theta).$$

$$\hat{Y} = W_L^T (\sigma(\dots \sigma(W_1^T \cdot D + b_1) \dots) + b_{L-1}) + b_L,$$

V_D : 유전자의 d 차원 전사 반응 기반의 약물 임베딩 벡터, $V_D \in \mathbb{R}^d$

V_C : 유전자의 d'차원 세포주 임베딩 벡터, $V_C \in \mathbb{R}^{d'}$

M: 약물 복용 정보

ΔT : 약물 투여 기간 정보

L: 은닉층의 수

w'_i : i번째 계층의 활성화 함수

b'_i : 가중치 행렬

σ : 바이어스

청구항 17

제 12 항에 있어서,

상기 d) 단계는,

단일 타겟인 경우, 상기 단일 타겟에 해당하는 세포주와 적어도 하나 이상의 대상 약물로 구성된 복수의 입력쌍에 대한 유전체 발현 반응 정보를 예측하고,

상기 e) 단계는,

상기 예측된 복수의 입력쌍에 대한 유전체 발현 반응 정보 중에서 기설정된 유전체 발현 패턴에 해당되는 입력쌍의 유전체 발현 반응 정보를 선택한 후 상기 선택된 입력쌍의 대상 약물을 단일 타겟의 맞춤형 약물로 선정하는 것인, 신약 후보 물질의 발굴 방법.

청구항 18

제 12 항에 있어서,

상기 d) 단계는,

유전자 발현의 억제가 필요한 복수의 타겟과 유전자 발현의 비억제가 필요한 복수의 비타겟을 설정하고, 상기 설정된 타겟 또는 비타겟에 해당하는 세포주와 적어도 하나 이상의 대상 약물로 구성된 복수의 입력쌍에 대한 유전체 발현 반응 정보를 예측하고,

상기 e) 단계는,

상기 예측된 복수의 입력쌍에 대한 유전체 발현 반응 정보 중에서 상기 설정된 타겟에 대한 유전체 발현 억제 정보와 상기 설정된 비타겟에 대한 유전체 발현 반응 정보를 모두 만족하는 적어도 하나 이상의 대상 약물을 선정하여 후보 약물그룹을 선정하는 것인, 신약 후보 물질의 발굴 방법.

청구항 19

제 12 항에 있어서,

상기 제1 데이터베이스에 기초한 약물과 세포주로 입력쌍을 구성한 경우, 해당 입력쌍은 이전의 약물 투입에 의한 이전의 세포주의 유전체 발현 반응 정보를 학습하기 위한 입력 데이터에 해당되고,

상기 제1 데이터베이스에 기초한 세포주와 제2 데이터베이스에 기초한 약물로 입력쌍을 구성한 경우, 해당 입력쌍은 새로운 약물 투입으로 인한 이전의 세포주의 유전체 발현 반응 정보를 확인하기 위한 입력 데이터에 해당되고,

상기 제1 데이터베이스에 기초한 약물과 제2 데이터베이스에 기초한 세포주로 입력쌍을 구성한 경우, 해당 입력쌍은 상기 이전의 약물 투입으로 인한 새로운 세포주의 유전체 발현 반응 정보를 예측하기 위한 입력 데이터에 해당되고,

상기 제2 데이터베이스에 기초한 약물과 세포주로 입력쌍을 구성한 경우, 해당 입력쌍은 새로운 약물 투입으로 인한 새로운 세포주의 유전자 발현 정보를 예측하기 위한 입력 데이터에 해당되는 것인, 신약 후보 물질의 발굴 방법.

청구항 20

기계 학습에 기반하여 신약 후보 물질을 발굴하기 위한 방법을 수행하기 위한 프로그램이 기록된 메모리; 및

상기 프로그램을 실행하기 위한 프로세서를 포함하며,

상기 프로세서는, 상기 프로그램의 실행에 의해,

복수 개의 약물 또는 세포주(Cell Line)에 대한 유전자 발현 정보, 유전자 발현 억제 정보, 각 약물의 구조 정보, 서로 다른 약물들 간의 약물 반응 유사성 정보를 포함하는 제1 데이터베이스를 제공하고,

상기 제1 데이터베이스에 기초하여 약물과 세포주로 구성된 입력쌍이 입력되면, 상기 입력쌍에 대해 상기 약물의 투입으로 인한 상기 세포주의 유전체 발현 반응 정보를 학습하는 유전체 반응 예측 모델을 제공하되,

상기 유전체 반응 예측 모델은, 상기 약물의 구조 정보에 대한 약물 임베딩 벡터, 상기 세포주의 유전체 정보에 대한 세포주 임베딩 벡터를 각각 산출한 후 상기 약물 임베딩 벡터와 세포주 임베딩 벡터를 이용하여 상기 약물과 세포주에 대해 후처리된 유전체 발현 반응 정보를 학습하는 것인, 신약 후보 물질의 발굴 시스템.

청구항 21

제 20 항에 있어서,

상기 프로세서는,

상기 유전체 발현 정보 또는 약물 반응 정보에 대해 알려지지 않은 약물 또는 세포주에 대한 특징 정보를 수집 및 관리하는 제2 데이터베이스를 제공하고,

상기 제2 데이터베이스에 기초한 약물 또는 세포주가 적어도 하나 이상이 포함되도록 입력쌍을 구성하고, 상기 입력쌍을 상기 유전체 반응 예측 모델에 입력하여 상기 약물의 투입으로 인한 세포주의 유전체 발현 반응 정보를 예측하고,

상기 예측된 유전체 발현 반응 정보가 사용자에 의해 설정된 유전체 발현 패턴에 해당되는 경우, 상기 제2 데이터베이스에 기초한 약물 또는 세포주를 상기 사용자가 원하는 약물 활동을 유도하기 위한 후보 물질로 선정하는 것인, 신약 후보 물질의 발굴 시스템.

청구항 22

제 21 항에 있어서,

상기 유전체 반응 예측 모델은,

상기 입력쌍 또는 후보 입력쌍이 입력되면 상기 약물의 화학 구조의 특징을 나타내는 구조 기반 벡터와 상기 세포주에 대한 유전체 발현 정보, 약물 투입 정보가 입력되는 입력층;

상기 약물의 구조 기반 벡터에 기초하여 약물 임베딩 벡터를 산출하는 제1 네트워크와 상기 세포주의 유전체 발현 정보에 기초하여 세포주 임베딩 벡터를 산출하는 제2 네트워크를 포함하는 임베딩층; 및

상기 약물 투입 정보, 상기 약물 임베딩 벡터, 상기 세포주 임베딩 벡터를 결합하여 상기 약물 투입 후의 유전체 발현 반응 정보를 출력하는 예측층을 포함하는 신약 후보 물질의 발굴 시스템.

청구항 23

제 22 항에 있어서,

상기 약물 투입 정보는 상기 약물의 복용 정보 또는 상기 약물 투여 기간 정보를 포함하는 것인, 신약 후보 물질의 발굴 시스템.

발명의 설명

기술 분야

[0001] 본 발명은 신약 후보 물질을 발굴하고 정밀 의료 시스템을 구현하기 위한 신약 후보 물질의 발굴 시스템 및 방법에 관한 것이다.

배경 기술

[0002] 유전자 발현 프로파일은 세포의 생물학적 상태를 나타내며, 전사 레벨(Transcriptional level)에서 약물의 생물학적 효과를 측정하는데 사용될 수 있다. 이러한 유전자 발현 프로파일은 약물 발견에 중요한 역할을 수행하기 때문에 HTS(High-Throughput Screening)는 유전자 발현에 대한 신속한 프로파일링, 원하는 활동을 유도하는 화합물 식별 등을 통해 유전자 발현 기반 약물 발견을 가속화할 수 있다.

[0003] 2006년 미국 브로드 연구소에서 공개한 CMap(Connectivity Map) 데이터세트는 다양한 화학유전체 데이터를 제공하고 있어, 약물의 기전해석, 미지의 약물 타겟 규명, 약물 용도 변경, 행동 메커니즘 예측, 개인화된 의약품, 질병의 생물학적 이해, 다수의 신약 재창출 및 분화 제어 약물 발굴 등에 활용되고 있다. CMap 데이터세트는 HTS가 입수한 약물 유도 유전자 발현 프로파일을 이용해 약물과 질병, 또는 두 가지 약물을 연결하는 개념에 기초하고 있다. CMap 데이터세트는 77개의 세포주(cell line)가 처리된 shRNA와 화합물과 같이 약 30,000개의 교란원(perturbagen)들의 130만개 유전자 발현 프로파일들을 포함한다.

[0004] 이와 같이, CMap 데이터세트는 신약 발굴에 유용하지만, 2만 개의 화합물 중에서 전체 8% 정도의 화합물 쌍을

실험하고 77개의 세포주만 실험되고 있다는 희소성과 같은 한계 사항이 있다. CMap 데이터세트의 희소성은 대규모 분석과 세포 매칭 비교를 실시하는 것을 어렵게 한다.

[0005] 한편, CMap에는 두 약물의 전사 대응 유사성을 나타내는 CMap 점수가 포함되어 있다. 화합물 쌍의 약물 반응 유사성을 얻기 위해 Cmap의 참조 데이터베이스인 터치스톤 데이터세트(Touchstone V1.0)를 활용하면, 2,428개의 화합물 쌍 중 9개 핵심 세포주(A375, A549, HA1E, HCC515, HEPG2, HT29, MCF7, PC3, VCAP)에서 검사되는 2400개의 CMap점수가 나와 있다. 이와 같이, CMap 데이터세트의 희소성으로 인해 사용자들은 터치스톤 데이터세트의 약물 목록에 포함되지 않은 다른 약물에 대한 CMap 점수를 얻을 수 없다는 문제점이 있다.

[0006] CMap 데이터세트는 방대한 양의 데이터를 포함하고 있음에도 불구하고, 적용 범위가 제한적이라는 한계가 있으므로, 다양한 세포주에서 새로운 약물을 테스트하여 약물의 작용 메커니즘을 파악하거나 새로운 세포주에서 다양한 약물을 테스트하여 세포주 특유의 바람직한 반응을 유도하는 약물 후보의 발굴 방법을 모색하는 것이 중요하다. 이를 위해, 신약 및 새로운 세포주로 많은 양의 새로운 샘플을 수집하는 것이 바람직하지만, 이는 엄청난 비용이 소요된다는 단점이 있다. 따라서, 약물 유도 유전자 발현 프로파일을 예측하고, HTS를 보완할 수 있는 정확한 예측 모델이 필요하다.

[0007] 약물에 의한 유전자 발현을 예측한 연구들 중에서는 물질에 대한 활동 스펙트라 예측 방법(Prediction of Activity Spectra for Substances, PASS), 약물 이웃 프로파일 예측 방법(Drug Neighbor Profile Prediction, DNPP), 고속 저층 텐서 완성 방법(Fast Low-Rank Tensor Completion, FaLRTC) 등이 있다.

[0008] PASS 방법은 약물의 하부 구조를 이용한 유전자의 활동 확률이나 불확실 가능성을 예측하는 베이지안 기반 모델이지만 치료 후 유전자 발현 수준을 억제할 수 없다는 문제점이 있다. DNPP 방법은 효과가 미미한 다른 약물의 다른 유전자 발현을 평균화하는 귀속 방법이고, FaLRTC 방법은 세포주, 약물, 유전자로 구성된 3D 텐서의 저층 행렬 근사치를 통해 텐서 완료를 수행하는 방법이다. DNPP와 FaLRTC는 희소한 텐서 채우기에 적합한 방법이지만, 신약이나 학습 세트에 없는 신약이나 새로운 세포주의 차등 유전자 발현 정도를 예측할 수 없고, 이전의 모든 모델들은 약물 복용량이나 약물 투여 기간을 전혀 고려하지 않고 있다는 문제점이 있다.

[0009] 이와 같이, 종래의 신약 개발 후보 물질 선정 과정은 화학적인 구조를 기반으로 단백질 등의 타겟과 결합 가능성 등을 고려하여 이루어지지만, 후보물질을 검증하는 과정(전임상 및 임상)에서 99.98%의 후보 물질이 탈락하고, 극소수의 화합물만 판매 단계까지 도달할 수 있다. 이는 체내의 복잡한 반응으로 인해 화학적 구조만을 고려한 후보 물질들이 탈락하기 때문이다. .

[0010] 또한, 종래에 약물 처방 방법은 질병에 관련된 타겟 물질에 반응하는 것으로 알려진 화합물을 처방한다. 즉 동일한 질병과 증상인 경우, 대부분 동일한 약물을 처방하게 된다. 그러나 약물에 대한 반응성은 개인의 유전체 발현 반응에 따라 다를 수 있다. 개인의 유전적 특성을 고려하지 않기 때문에 약효가 기대치와 다른 경우가 많으므로, 약물의 유전체 발현 반응을 예측하여 정밀한 약물 처방을 도와줄 수 있는 기술이 요구되고 있다.

선행기술문헌

[0011] 대한민국 등록특허 제 10-1870963 호(발명의 명칭 : 테트라스파닌-2를 이용한 당뇨병의 예방 또는 치료용 조성물과 당뇨병 치료제 스크리닝 방법)

발명의 내용

해결하려는 과제

[0012] 본 발명은 전술한 문제점을 해결하기 위하여, 본 발명의 일 실시예에 따라 약물의 화학 구조 정보와 약물 투입으로 인한 생체 반응을 고려하여 신약 후보 물질을 발굴하고, 환자 개인에 대한 약물 투입 후의 유전체 발현 반응을 예측하여 정밀한 약물 처방이 가능해질 수 있는 신약 후보 물질의 발굴 시스템 및 그 방법을 제공하는 것에 목적이 있다.

[0013] 다만, 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

[0014] 상기한 기술적 과제를 달성하기 위한 기술적 수단으로서 본 발명의 일 실시예에 따른, 신약 후보 물질의 발굴

방법은, 기계 학습에 기반하여 신약 후보 물질을 발굴하기 위한 신약 후보 물질의 발굴 시스템에 의해 수행되는 신약 후보 물질의 발굴 방법에 있어서, a) 복수 개의 약물 또는 세포주(Cell Line)에 대한 유전자 발현 정보, 유전자 발현 억제 정보, 각 약물의 구조 정보, 서로 다른 약물들 간의 약물 반응 유사성 정보를 포함하는 데이터베이스를 제공하는 단계; 및 b) 상기 데이터베이스에 기초하여 약물과 세포주로 구성된 입력쌍이 입력되면, 상기 입력쌍에 대해 상기 약물의 투입으로 인한 상기 세포주의 유전체 발현 반응 정보를 학습하는 유전체 반응 예측 모델을 제공하는 단계를 포함하되, 상기 유전체 반응 예측 모델은, 상기 약물의 구조 정보에 대한 약물 임베딩 벡터, 상기 세포주의 유전체 정보에 대한 세포주 임베딩 벡터를 각각 산출한 후 상기 약물 임베딩 벡터와 세포주 임베딩 벡터를 이용하여 상기 약물과 세포주에 대해 후처리된 유전체 발현 반응 정보를 학습하는 것이다.

[0015] 또한, 본 발명의 다른 일 실시예에 따른 신약 후보 물질의 발굴 방법은, 기계 학습에 기반하여 신약 후보 물질을 발굴하기 위한 신약 후보 물질의 발굴 시스템에 의해 수행되는 신약 후보 물질의 발굴 방법에 있어서, a) 복수 개의 약물 또는 세포주(Cell Line)에 대한 유전자 발현 정보, 유전자 발현 억제 정보, 각 약물의 구조 정보, 서로 다른 약물들 간의 약물 반응 유사성 정보를 포함하는 제1 데이터베이스를 제공하는 단계; b) 상기 제1 데이터베이스에 기초하여 약물과 세포주로 구성된 입력쌍이 입력되면, 유전체 반응 예측 모델을 통해 상기 입력쌍에 대한 상기 약물의 투입으로 인한 상기 세포주의 유전체 발현 반응 정보를 학습하는 단계; c) 상기 유전체 발현 정보 또는 약물 반응 정보에 대해 알려지지 않은 약물 또는 세포주에 대한 특징 정보를 수집 및 관리하는 제2 데이터베이스를 제공하는 단계; d) 상기 제2 데이터베이스에 기초한 약물 또는 세포주가 적어도 하나 이상이 포함되도록 입력쌍을 구성하고, 상기 입력쌍을 상기 유전체 반응 예측 모델에 입력하여 상기 약물의 투입으로 인한 세포주의 유전체 발현 반응 정보를 예측하는 단계; 및 e) 상기 예측된 유전체 발현 반응 정보가 사용자에게 의해 설정된 유전체 발현 패턴에 해당되는 경우, 상기 제2 데이터베이스에 기초한 약물 또는 세포주를 상기 사용자가 원하는 약물 활동을 유도하기 위한 후보 물질로 선정하는 단계를 포함하되, 상기 유전체 반응 예측 모델은, 상기 약물의 구조 정보에 대한 약물 임베딩 벡터, 상기 세포주의 유전체 정보에 대한 세포주 임베딩 벡터를 각각 산출한 후 상기 약물 임베딩 벡터와 세포주 임베딩 벡터를 이용하여 상기 약물과 세포주에 대해 후처리된 유전체 발현 반응 정보를 학습 및 예측하는 것이다.

[0016] 본 발명의 일 실시예에 따른 신약 후보 물질의 발굴 시스템은, 기계 학습에 기반하여 신약 후보 물질을 발굴하기 위한 방법을 수행하기 위한 프로그램이 기록된 메모리; 및 상기 프로그램을 실행하기 위한 프로세서를 포함하며, 상기 프로세서는, 상기 프로그램의 실행에 의해, 복수 개의 약물 또는 세포주(Cell Line)에 대한 유전자 발현 정보, 유전자 발현 억제 정보, 각 약물의 구조 정보, 서로 다른 약물들 간의 약물 반응 유사성 정보를 포함하는 제1 데이터베이스를 제공하고, 상기 제1 데이터베이스에 기초하여 약물과 세포주로 구성된 입력쌍이 입력되면, 상기 입력쌍에 대해 상기 약물의 투입으로 인한 상기 세포주의 유전체 발현 반응 정보를 학습하는 유전체 반응 예측 모델을 제공하되, 상기 유전체 반응 예측 모델은, 상기 약물의 구조 정보에 대한 약물 임베딩 벡터, 상기 세포주의 유전체 정보에 대한 세포주 임베딩 벡터를 각각 산출한 후 상기 약물 임베딩 벡터와 세포주 임베딩 벡터를 이용하여 상기 약물과 세포주에 대해 후처리된 유전체 발현 반응 정보를 학습하는 것이다.

발명의 효과

[0017] 전술한 본 발명의 과제 해결 수단에 의하면, 본 발명은 체내의 복잡한 생체 반응을 고려하여 약물의 화학 구조, 유전체 발현 정보, 약물 복용 정보, 약물 투여 기간 정보를 모두 결합하여 사용자가 원하는 유전체 발현 패턴을 갖는 신약 후보 물질을 발굴할 수 있다.

[0018] 또한, 본 발명은 유전자 발현을 억제하고자 하는 타겟과 유전체 발현을 억제하지 않아야 하는 비타겟들의 약물 투입 후의 유전체 발현 반응을 동시에 예측할 수 있어, 여러 타겟들과 비타겟과의 상호 작용을 검토하여 후보 물질의 발굴에 대한 정확도가 크게 향상될 수 있다.

[0019] 본 발명은 환자 개인의 유전적 특성을 고려하여 여러 대상 약물 중 환자의 유전체 발현 반응 패턴과 가장 가까운 약물을 예측하여 해당 약물을 환자 맞춤형 약물로 처방함으로써 약효 기대치를 높일 수 있고, 암과 같은 유전 질환을 위한 정밀 의료에 활용될 수 있다.

도면의 간단한 설명

[0020] 도 1은 본 발명의 일 실시예에 따른 신약 후보 물질의 발굴 시스템의 구성을 나타낸 도면이다.

도 2는 본 발명의 일 실시예에 따른 신약 후보 물질의 발굴 방법을 설명하는 순서도이다.

도 3은 본 발명의 일 실시예에 따른 유전체 반응 예측 모델을 설명하는 도면이다.

도 4는 본 발명의 다른 일 실시예에 따른 신약 후보 물질의 발굴 방법을 설명하는 순서도이다.

도 5는 본 발명의 일 실시예에 따른 신약 후보 물질의 발굴 방법에서 사용된 입력 데이터를 설명하는 도면이다.

도 6은 본 발명의 일 실시예에 따른 신약 후보 물질의 발굴 방법의 파이프라인 개요를 설명하는 도면이다.

도 7은 본 발명의 일 실시예에 따라 4개의 테스트용 데이터 세트를 이용한 유전체 반응 예측 모델을 포함한 기계 학습 기반 베이스 라인 모델의 성능을 나타낸 도면이다.

도 8은 본 발명의 일 실시예에 따라 플뤼페나진의 임베딩 벡터와 유사한 임베딩 벡터를 내장한 ZINC15 화합물의 상위 10개 후보 물질들을 열거한 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0021] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.
- [0022] 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. 또한 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미하며, 하나 또는 그 이상의 다른 특징이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0023] 이하의 실시예는 본 발명의 이해를 돕기 위한 상세한 설명이며, 본 발명의 권리 범위를 제한하는 것이 아니다. 따라서 본 발명과 동일한 기능을 수행하는 동일 범위의 발명 역시 본 발명의 권리 범위에 속할 것이다.
- [0025] 이하 첨부된 도면을 참고하여 본 발명의 일 실시예를 상세히 설명하기로 한다.
- [0026] 도 1은 본 발명의 일 실시예에 따른 신약 후보 물질의 발굴 시스템의 구성을 나타낸 도면이다.
- [0027] 도 1을 참조하면, 신약 후보 물질의 발굴 시스템(100)은 통신 모듈(110), 메모리(120), 프로세서(130) 및 데이터베이스(140)를 포함한다.
- [0028] 상세히, 통신 모듈(110)은 통신망(300)과 연동하여 신약후보 물질의 발굴 시스템(100)으로 송수신되는 신호를 패킷 데이터 형태로 제공하는 데 필요한 통신 인터페이스를 제공한다. 여기서, 통신 모듈(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.
- [0029] 메모리(120)는 기계 학습에 기반하여 신약 후보 물질의 발굴 방법을 수행하기 위한 프로그램이 기록된다. 또한, 프로세서(130)가 처리하는 데이터를 일시적 또는 영구적으로 저장하는 기능을 수행한다. 여기서, 메모리(120)는 휘발성 저장 매체(volatile storage media) 또는 비휘발성 저장 매체(non-volatile storage media)를 포함할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.
- [0030] 프로세서(130)는 기계 학습에 기반하여 신약 후보 물질의 발굴 방법을 제공하는 전체 과정을 제어한다. 프로세서(130)는 세포주가 약물에 의한 유전체 발현 반응을 예측하기 위해 유전체 반응 예측 모델을 적용한다. 이때, 유전체 반응 예측 모델은 약물의 구조 정보, 세포주의 전처리된 유전자 발현 정보, 약물 복용 정보, 약물 투여 기간 정보를 입력 데이터로 사용하며, 약물 투입 후의 세포주에 대한 유전체 발현 반응 정보를 출력한다. 이러한 프로세서(130)가 수행하는 각각의 동작에 대해서는 추후 보다 상세히 살펴보기로 한다.
- [0031] 여기서, 프로세서(130)는 프로세서(processor)와 같이 데이터를 처리할 수 있는 모든 종류의 장치를 포함할 수 있다. 여기서, '프로세서(processor)'는, 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.

- [0032] 데이터베이스(140)는 신약 후보 물질의 발굴방법을 수행하면서 누적되는 데이터가 저장된다. 예컨대, 이미 알려진 화합물 데이터 집합을 제1 데이터베이스(141)에 저장하고, 제1 데이터베이스(141)에 저장된 데이터 집합을 제외한 화합물 데이터 집합을 제2 데이터베이스(142)에 저장할 수 있다. 여기서, 화합물은 이미 알려진 약물, 신약 후보 물질 등이 될 수 있다.
- [0033] 제1 데이터베이스(141)에는 약물의 유전자 발현 정보, 유전자 발현 정보로부터 예측된 차등 발현 유전자(differentially expressed gene, DEG)의 발현량을 포함한 약물 유도 전사체, 서로 다른 약물들 간의 약물 반응 유사성 정보인 CMap 점수를 포함하는 연결 지도(Connectivity Map, CMap) 데이터셋을 구성한다. 이때, 약물 반응 유사성 정보는 화합물들 사이의 CMap 점수가 될 수 있다. CMap 점수는 화합물, 9개 핵심 세포 라인(A375, A549, HA1E, HCC515, HEPG2, HT29, MCF7, PC3, VCAP)에 걸친 유전자의 기능 이득, 유전자의 기능 손실과 같은 섭동학자의 프로파일된 특징에 기초한 유사도 점수이다.
- [0034] 이때, 약물 유도 전사체 데이터는 기존의 약물-타겟 및 생리활성 데이터와 차별화된 정보를 제공한다. 즉 약물 유도 전사체 데이터는 약물의 알려진 온타겟(on-target) 뿐만 아니라 미지의 오프 타겟(off-target)에 의한 효과까지도 반영하며, 전사체 수준에서 약물 작용 기전(mode of action)에 대한 객관적이고 총체적인 정보를 제공하고, 질환 모델 전사체와의 비교 분석을 통해 신약 재창출이 가능하도록 한다.
- [0035] 따라서, CMap을 활용한 제1 데이터베이스(141)는 기존의 단일 타겟형 신약 후보 물질 선정을 보완하여 약물 투입으로 인한 세포주의 유전체 발현 반응 정보를 이용하여 신약 후보 물질을 발굴하도록 할 수 있다.
- [0036] 제2 데이터베이스(142)는 ZINC15 데이터베이스가 될 수 있고, ZINC15 데이터베이스는 2억 3천만 개 이상의 화합물을 포함하고 있으며, 각 화합물은 고유한 등록번호(registry number)와 이름이 부여되고, 구매 가능하며, 리핀스키(Lipinski) 규칙을 충족하는 약 1만 6천 개의 ZINC15 화합물을 선정할 수 있다. 이렇게 선정된 화합물들은 제1 데이터베이스(141)에 등록되지 않은, 즉 유전자 발현 정보가 알려지지 않은 화합물이다.
- [0037] 도 2는 본 발명의 일 실시예에 따른 신약 후보 물질의 발굴 방법을 설명하는 순서도이고, 도 3은 본 발명의 일 실시예에 따른 유전체 반응 예측 모델을 설명하는 도면이다.
- [0038] 도 2를 참조하면, 신약 후보 물질의 발굴 방법은, 복수 개의 약물 또는 세포주(Cell Line)에 대한 유전자 발현 정보, 유전자 발현 억제 정보, 각 약물의 구조 정보, 서로 다른 약물들 간의 약물 반응 유사성 정보를 포함하는 제1 데이터베이스(141)를 제공한다(S110).
- [0039] 프로세서(130)는 약물의 화학 구조와 세포주의 유전체 발현 반응 정보를 통해 해당 약물로 인한 생체 반응을 고려하여 후보 물질을 발굴하기 위한 유전체 반응 예측 모델(300)을 제공한다(S120).
- [0040] 도 3에 도시된 바와 같이, 유전체 반응 예측 모델(300)은 다차원의 다층 신경망(Multi-Layer Perceptron, MLP)으로 구성되고, 입력층(310), 임베딩층(320) 및 예측층(330)을 포함한다.
- [0041] 입력층(310)은 입력쌍이 입력되면 약물의 화학 구조의 특징을 나타내는 구조 기반 벡터, 세포주에 대한 유전체 발현 정보, 약물 복용 정보, 약물 투여 기간 정보가 입력된다.
- [0042] 임베딩층(320)은 약물의 구조 기반 벡터에 기초하여 약물 임베딩 벡터를 산출하고, 세포주의 유전체 발현 정보에 기초하여 세포주 임베딩 벡터를 산출한다. 이때, 임베딩층(320)은 약물 임베딩 벡터를 산출하는 제1 네트워크(MLP_D) 및 세포주 임베딩 벡터를 산출하는 제2 네트워크(MLP_C)를 포함한다.
- [0043] 예측층(330)은 약물 복용 정보, 약물 투여 기간 정보, 약물 임베딩 벡터, 세포주 임베딩 벡터를 결합하여 약물 투입 후의 세포주에 대한 유전체 발현 반응 정보를 출력한다.
- [0044] 유전체 반응 예측 모델(300)은 제1 데이터베이스(141)에 기초하여 약물과 세포주로 구성된 입력쌍이 입력되면(S130), 약물의 구조 정보에 대한 약물 임베딩 벡터와 세포주의 유전체 정보에 대한 세포주 임베딩 벡터를 각각 산출한 후 약물 임베딩 벡터와 세포주 임베딩 벡터를 이용하여 약물과 세포주에 대해 후처리된 유전체 발현 반응 정보를 학습한다(S140, S150). 이때, 유전체 반응 예측 모델(300)은 약물 임베딩 벡터와 세포주 임베딩 벡터의 유사도를 계산하고, 계산된 유사도를 이용해 예측층(330)에서 예측된 유전체 발현 반응 정보가 입력쌍의 약물 반응 유사성 정보가 되도록 학습한다.

- [0045] 세포주 임베딩 벡터는 세포주의 전처리된 유전자 발현 정보와 기 학습된 유전체 반응 예측 모델(300)을 이용해 약물 반응과 연관된 벡터로 임베딩되고, 약물 임베딩 벡터는 약물의 화학 구조의 특징을 나타내는 구조 기반 벡터와 기 학습된 유전체 반응 예측 모델(300)을 이용해 세포주의 유전체 발현 반응과 연관된 벡터로 임베딩된다.
- [0046] 약물 임베딩 벡터는(V_D) d' 차원의 전사적 반응 기반의 약물을 훈련하기 위한 제1 네트워크(MLP_D)에서 하기 수학식 1에 의해 생성되고, 세포주 임베딩 벡터(V_C)는 전사적 반응 기반의 세포주를 훈련하기 위한 제2 네트워크(MLP_C)에서 하기 수학식 2에 의해 생성된다. 이때, $C=\{C_1, C_2, \dots, C_{70}\}$ 는 세포주의 데이터 집합이고, $D=\{D_1, D_2, \dots, D_{20336}\}$ 는 약물의 데이터 집합이며, 각 세포주 $C \in R^{978}$ 은 유전자의 978차원의 고차원 발현 데이터로 표현되고, 약물은 화학구조식 파일을 소프트웨어 도구 세트(RDkit, RDkit3, OpenBabel, Marvin View의 molconvert 등)를 이용하여 ECFP(Extended Connectivity FingerPrint)로 표현된다. ECFP는 화합물의 바이너리 벡터를 생성하여 화합물이 특정 하부구조를 가지고 있는지 여부를 나타낸다. 약물은 2048 차원의 ECFP 로 표현 된다.
- [0047] 또는 약물 임베딩 벡터는 ECFP의 일차원 벡터 표기 방식 외에도, SMILES(Simplified Molecular Input Line Entry System), WLN(Wiswesser line notation), InChI(International Chemical Identifier)를 포함한 화합물의 문자열 표기 방식 중 어느 하나의 방식을 이용하여 각 약물의 화합물의 구조를 문자열로 표기하는 문자열 표기 방식, 또는 Graph Neural Networks(GNNs), 접합 트리(Junction Tree) 알고리즘 등을 이용하여 인접행렬과 노드 특징(feature)을 갖는 그래프 구조 표기 방식으로 표현될 수 있다.
- [0048] SMILES을 이용한 문자열 표기 방식은 WLN, InChI, ROSDAL 등과 마찬가지로 화합물의 구조를 문자열로 나타내는 방법 중 하나로서, 쉽게 화합물질의 구조를 문자열로 나타낼 수 있어 여러 분야에서 활용되고 있다.
- [0049] 이러한 SMILES을 이용한 문자열 표기 방식은 크게 6개의 규칙만 적용하는 것으로 화학 구조를 문자열로 바꿀 수 있다. SMILES의 규칙을 살펴보면, ① 원자는 표준 원소기호로 나타내고, ② 수소 원자는 가능한 모든 곳에 연결되어 있다고 하여 표기를 생략하며, ③ 이웃한 원자는 바로 인접해서 표기하고, ④ 2중결합은 '=' 으로, 3중결합은 '#' 으로 나타내며, ⑤ 가지는 괄호 '(' 를 써서 나타내고, ⑥ 고리는 고리를 만드는 두 원자에 숫자를 표기하는 것으로 나타내는데, 방향족 고리는 원자를 소문자로 표시한다.
- [0050] 한편, 임베딩 벡터의 그래프 구조 표기 방식은 인접 행렬과 노드 특징 행렬을 GNNs의 입력 정보로 입력하는데, 인접 행렬은 각 노드들의 연결 정보를 포함하고 있어 결합 종류에 대한 정보를 반영하고, 노드 특징은 각 원자의 종류와 원자가 등의 정보를 반영한다. 따라서, 약물 임베딩 벡터는 각 약물의 구조 정보를 활용하여 GNNs를 통해 인접 행렬과 노드 특징을 갖는 그래프 구조로 표현될 수 있다.
- [0051] [수학식 1]
- [0052]
$$V_D = \overline{W}_L^T (\sigma(\dots \sigma(\overline{W}_1^T \cdot D + \overline{b}_1) \dots) + \overline{b}_{L-1}) + \overline{b}_L,$$
- [0053] [수학식 2]
- [0054]
$$V_C = W_L'^T (\sigma(\dots \sigma(W_1'^T \cdot C + b'_1) \dots) + b'_{L-1}) + b'_L,$$
- [0055] 수학식 1 및 수학식 2에서, V_D 는 유전자의 d 차원 전사 반응 기반의 약물 임베딩 벡터($V_D \in \mathbb{R}^d$)이고, V_C 는 유전자
- $\overline{W}_i^T$
- 의 d' 차원 세포주 임베딩 벡터($V_C \in \mathbb{R}^{d'}$)이며, L 은 은닉층의 수, $\overline{W}_i^T$ 는 제1 네트워크의 i 번째 계층의 활성화
- $\overline{b}_i^T$
- 함수, $\overline{b}_i^T$ 는 제1 네트워크의 가중치 행렬, $W_i'^T$ 는 제2 네트워크의 i 번째 계층의 활성화 함수, $b_i'^T$ 는 제2 네트워크의 가중치 행렬, σ 는 바이어스를 각각 나타낸다.
- [0056] 이러한 유전체 반응 예측 모델(300)은 다차원의 다층 신경망(Multi-Layer Perceptron, MLP)으로 구성되어 있어, 입력층(310)은 하기 수학식 3에 의한 입력 데이터를 임베딩층(320)으로 입력하고, 예측층(330)은 하기 수학식 4
- $\hat{Y}$
- 에 의해 예측된 유전체 발현 반응 정보를 출력 데이터($\hat{Y}$)로 출력한다.

[0057] [수학식 3]

$$X = [V_C; V_D; M; \Delta T].$$

[0058]

[0059] [수학식 4]

$$\hat{Y} = f(C, D, M, \Delta T | \theta).$$

[0060]

$$\hat{Y} = W_L^T (\sigma(\dots \sigma(W_1^T \cdot D + b_1) \dots) + b_{L-1}) + b_L,$$

[0061]

[0062] 수학식 3 및 수학식 4에서, M은 약물 복용 정보($M \in \mathbb{R}$), ΔT 는 약물 투여 기간 정보($\Delta T \in \mathbb{R}$)를 각각 나타낸다.

특히 수학식 4에서, $\hat{Y} \in \mathbb{R}^{978}$ 이고, θ 는 유전체 반응 예측 모델(300)의 파라미터이다.

[0063] 유전체 반응 예측 모델(300)은 비용 함수를 하기 수학식 5에 의한 평균 제곱 오차(Mean square error, MSE)를 이용해 계산하고, 비용 함수의 결과값이 최소화되도록 학습을 통해 최적의 가중치를 결정한다.

[0064] [수학식 5]

$$MSE = \frac{1}{N} \sum_{i=1}^N \|Y_i - \hat{Y}_i\|_2^2,$$

[0065]

[0066] 수학식 5에서, Y_i 는 i 번째 입력 데이터에 대한 실제 유전체 발현 반응 값, $\hat{Y}_i$ 는 i 번째 입력 데이터에 대한 예측된 유전체 발현 반응 값, N은 데이터베이스 내의 입력 데이터의 수를 각각 나타낸다.

[0067] 하기 표 1에는 유전체 반응 예측 모델의 각 기호에 대한 설명을 나타낸다.

[0068] [표 1]

Notation	Description
$C, \mathcal{C}$	Cell line and set of cell lines
$D, \mathcal{D}$	Drug and set of drugs
V_C, V_D	Cell and drug embedding vectors
W', b'	Weight matrix and bias term of MLP_C
$\bar{W}, \bar{b}$	Weight matrix and bias term of MLP_D
W, b	Weight matrix and bias term of prediction layer
$M, \Delta T$	Drug dosage and duration of drug administration
$Y, \hat{Y}$	True and predicted post-treatment profile

[0069]

[0070] 일례로, 확률적 경사 하강(Stochastic gradient descent) 알고리즘을 사용하여, 계산된 MSE를 최소화하여 유전체 반응 예측 모델(300)의 파라미터(θ)를 최적화하고, 아담(Adam, Adaptive Moment Estimation) 옵티마이저(Optimizer)를 적용하여 256의 배치 크기와 0.001의 학습률을 사용한다.

[0071] 도 4는 본 발명의 다른 일 실시예에 따른 신약 후보 물질의 발굴 방법을 설명하는 순서도이고, 도 5는 본 발명의 일 실시예에 따른 신약 후보 물질의 발굴 방법에서 사용된 입력 데이터를 설명하는 도면이다.

[0072] 도 4를 참조하면, 신약 후보 물질의 발굴 방법은, 유전체 발현 정보 또는 약물 반응 정보에 대해 알려지지 않은 약물 또는 세포주에 대한 특징 정보를 수집 및 관리하는 제2 데이터베이스(142)를 제공한다(S210).

[0073] 프로세서(130)는 제2 데이터베이스(142)에 기초한 약물 또는 세포주가 적어도 하나 이상이 포함되도록 입력쌍을 구성하고, 입력쌍을 유전체 반응 예측 모델(300)의 입력 데이터로 제공한다(S220).

[0074] 유전체 반응 예측 모델(300)은 입력쌍에 대해 약물의 투입으로 인한 세포주의 유전체 발현 반응 정보를 예측한 출력 데이터를 출력한다(S230).

- [0075] 프로세서(130)는 예측된 유전체 발현 반응 정보가 사용자에게 의해 설정된 유전체 발현 패턴에 해당되는 경우 (S240), 제2 데이터베이스(142)에 기초한 약물 또는 세포주를 사용자가 원하는 약물 활동을 유도하기 위한 후보 물질로 선정한다(S250).
- [0076] 프로세서(130)는 제1 데이터베이스(141)와 제2 데이터베이스(142)를 활용하여 입력쌍을 구성하여 유전체 반응 예측 모델(300)에 학습용 입력 데이터 70%, 유효성 검사용 입력 데이터 20%, 테스트용 입력 데이터 10%를 제공한다.
- [0077] 도 5에 도시된 바와 같이, 입력 데이터는 4개의 데이터셋으로 구성되는데, (a)(구 약물(Old drug), 구 세포주(Old cell line)), (b)((신 약물(New drug), 구 세포주), (c) (구 약물, 신 세포주(New cell line)), (d)(신 약물, 신 세포주)로 입력쌍을 구성할 수 있다.
- [0078] (구 약물, 구 세포주)의 데이터 세트는 학습용 입력 데이터로서, 제1 데이터베이스(141)에 기초한 약물과 세포주로 입력쌍을 구성한다.
- [0079] (신 약물, 구 세포주)의 데이터 세트는 제1 데이터베이스(141)에 기초한 세포주와 제2 데이터베이스(142)에 기초한 새로운 약물로 입력쌍을 구성하고, 해당 입력쌍은 약물의 후보 물질을 선정하기 위해 새로운 약물 투입으로 인한 구 세포주의 유전체 발현 반응 정보를 확인하기 위한 입력 데이터에 해당된다.
- [0080] (구 약물, 신 세포주)의 데이터 세트는 제1 데이터베이스(141)에 기초한 약물과 제2 데이터베이스(142)에 기초한 세포주로 입력쌍을 구성하고, 해당 입력쌍은 새로운 세포주에 미치는 약물의 영향을 파악하기 위한 입력 데이터에 해당된다.
- [0081] (신 약물, 신 세포주)의 데이터 세트는 제2 데이터베이스(142)에 기초한 약물과 세포주로 입력쌍을 구성하고, 해당 입력쌍은 새로운 약물 투입으로 인한 새로운 세포주의 유전자 발현 정보를 예측하기 위한 입력 데이터에 해당된다.
- [0082] 표 2는 사전처리된 데이터셋의 통계를 나타낸 것으로서, 학습용, 유효성 검사용, 테스트용의 데이터 세트를 요약한 것이다.

[0083] [표 2]

Statistics	Train	Validation	OO	ON	NO	NN
# of tuples	127,708	18,511	7,095	7,666	25,190	1,306
# of drugs	18,196	5,673	4,655	10,173	1,018	533
# of cells	57	70	57	56	6	6
Average # of tuples per cell	2,240.5	264.4	124.5	4,198.3	137.3	217.7
Average # of tuples per drug	7	3.2	1.5	2.5	7.6	2.4

OO, ON, NO, and NN denote the four test sets, which are (Old drug, Old cell line), (Old drug, New cell line), (New drug, Old cell line), and (New drug, New cell line), respectively.

- [0084]
- [0085] 도 6은 본 발명의 일 실시예에 따른 신약 후보 물질의 발굴 방법의 파이프라인 개요를 설명하는 도면이다.
- [0086] 본 발명의 신약 후보 물질의 발굴 방법은 약물 후보물질들의 유전체 발현 반응을 예측하여 원하는 발현 패턴을 주는 후보물질을 발굴하는 신약 후보 물질의 발굴 시스템(100) 및 환자에게 어떤 약물을 투약해야 이상적인 효과를 볼 수 있는지 예측하는 정밀 의료 시스템(400)에 적용될 수 있다.
- [0087] 신약 후보 물질의 발굴 시스템(100)은 유전자 발현의 억제가 필요한 복수의 타겟과 유전자 발현의 비억제가 필요한 복수의 비타겟을 설정하고, 설정된 타겟 또는 비타겟에 해당하는 세포주와 적어도 하나 이상의 대상 약물로 구성된 복수의 입력쌍에 대한 유전체 발현 반응 정보를 예측한다. 그 후에 예측된 복수의 입력쌍에 대한 유전체 발현 반응 정보 중에서 기설정된 타겟에 대한 유전체 발현 억제 정보와 기설정된 비타겟에 대한 유전체 발현 반응 정보를 모두 만족하는 적어도 하나 이상의 대상 약물을 선정하여 후보 약물그룹을 선정한다.
- [0088] 예를 들어, RET, BRAF, SRC, S6K 의 유전자 발현은 억제하고, 동시에 MKNK1, TTK, Erk8, PDK1, PAK3 유전자 발현은 억제하지 않는 후보 약물 그룹을 선정할 수 있다.
- [0089] 현재 신약 개발 파이프라인은 약물의 구조에 기반한 후보물질 발굴에 의존하고 있기 때문에 주로 단일타겟(예,

RET 하나만)을 기준으로 후보 물질의 발굴이 이루어지고 있다. 따라서 다른 타겟들 및 비타겟들과의 상호 작용은 검토가 어렵기 때문에 전임상 임상에서 실패 확률이 높아지는 단점이 있다. 그러나 신약 후보 물질의 발굴 시스템(100)은 유전체 반응 예측 모델(300)을 통해 여러 타겟들과 비타겟들의 투약 후 유전체 발현 반응을 동시에 예측할 수 있어 후보 물질의 발굴 정확도를 크게 향상시킬 수 있다.

[0090] 한편, 정밀 의료 시스템(400)은 단일 타겟인 경우, 단일 타겟에 해당하는 세포주와 적어도 하나 이상의 대상 약물로 구성된 복수의 입력쌍에 대한 유전체 발현 반응 정보를 예측하고, 예측된 복수의 입력쌍에 대한 유전체 발현 반응 정보 중에서 기설정된 유전체 발현 패턴에 해당되는 입력쌍의 유전체 발현 반응 정보를 선택한 후 선택된 입력쌍의 대상 약물을 단일 타겟의 맞춤형 약물로 선정할 수 있다.

[0091] 즉, 정밀 의료 시스템은 환자 개인의 유전체 발현 정보와 투약 대상 약물을 입력쌍으로 하여 유전체 반응 예측 모델에 입력하고, 유전체 반응 예측 모델에서 해당 환자의 대상 약물 투약 후 유전체 발현 반응을 예측한다. 따라서, 여러 대상 약물에 대해 예측한 환자의 유전체 발현 반응 패턴 중 가장 정상에 가까운 패턴을 나타내는 약물을 환자의 개인화 처방 약물로 추천할 수 있다. 따라서, 정밀 의료 시스템은 암과 같은 유전 질병을 위한 정밀 의료 및 맞춤 의약품 처방에 활용할 수 있다.

[0092] 한편, 유전체 반응 예측 모델을 TRNet(a deep neural network-based transcriptional response prediction model)이라고 할 경우, TRNet의 성능을 5개의 기계 학습 기반의 베이스라인 모델과 비교하여 평가할 수 있다.

[0093] 5개의 기계 학습 기반 베이스 라인 모델에는 무작위 포레스트 회귀 분석기(Random Forest regressor, RF), k-최근접 이웃 회귀 분석기(k-Nearest Neighbors regressor, kNN), 릿지 회귀 분석기(Ridge regressor), 엑스트라 트리 회귀 분석기(eXtra Tree regressor, XT) 및 MLP 회귀 분석기가 사용될 수 있다. 주어진 세포주의 전처리 유전자 발현 프로파일, 주어진 약물의 화학적 구조 기반의 벡터, 약물 복용 정보, 약물 투여 기간 정보를 입력 벡터로 사용하고, TRNet과 같이 유효성 검사용 입력 데이터를 이용하여 각 모델의 최적 하이퍼 파라미터를 찾는다.

[0094] 각 모델에 대해 고려된 하이퍼 파라미터는 표 3에 기재되어 있고, 최적의 하이퍼 파라미터를 사용하여 각 모델은 학습 및 검증용 입력 데이터에 대해 학습을 진행한다.

[0095] [표 3]

Model	Hyper-parameters	Range
TRNet	1st layer dimension of MLP_C	[400,800]
	$ V_C $	[200,400]
	1st layer dimension of MLP_D	[600,1200]
	$ V_D $	[200,400]
	1st layer dimension of prediction layer	[400,800]
	2nd layer dimension of prediction layer	[200,400]
RF	# of estimators	[128, 256]
	maximum features	["log2", "sqrt"]
	maximum depth	[10, 50]
XT	# of estimators	[128, 256]
	maximum features	["log2", "sqrt"]
	maximum depth	[10, 40]
KNN	K	[10, 90]
Ridge	C	[0.001, 10000]
MLP	# of layers	[1,3]
	1st layer dimension	[64,512]
	2nd layer dimension	[64,256]
	3rd layer dimension	[64,128]

[0096]

[0097] TRNet과 기계 학습 기반 베이스 라인 모델의 성능을 평가하기 위해서는 하기 수학적 6의 평가지표를 사용하여

실제 후처리 유전자 발현 반응 정보와 예측된 후처리 유전자 발현 반응 정보를 비교한다.

[수학식 6]

$$MAE = \frac{1}{N} \sum_{i=1}^N \|Y_i - \hat{Y}_i\|, MSE = \frac{1}{N} \sum_{i=1}^N \|Y_i - \hat{Y}_i\|_2^2.$$

즉, 코사인 척도(CD)를 유전자 발현 패턴간 유사도로 설정하고, 하기 수학식 7과 같이 계산한다.

[수학식 7]

$$CD = \frac{1}{N} \sum_{i=1}^N (1 - \frac{Y_i \cdot \hat{Y}_i}{\|Y_i\| \|\hat{Y}_i\|}),$$

피어슨 상관관계(PC)와 R-제곱(R²)은 하기 수학식 8과 같이 계산한다.

[수학식 8]

$$PC = \frac{1}{N} \sum_{i=1}^N \frac{\sum_{j=1}^{978} (Y_{i,j} - \bar{Y})(\hat{Y}_{i,j} - \bar{\hat{Y}})}{\sqrt{\sum_{j=1}^{978} (Y_{i,j} - \bar{Y})^2} \sqrt{\sum_{j=1}^{978} (\hat{Y}_{i,j} - \bar{\hat{Y}})^2}},$$

$$R^2 = \frac{1}{N} \sum_{i=1}^N (1 - \frac{\|Y_{i,j} - \hat{Y}_{i,j}\|_2^2}{\|Y_{i,j} - \bar{Y}\|_2^2}),$$

수학식 8에서, Y_i 는 i 번째 입력 데이터에 대한 실제 유전체 발현 반응 값의 평균값이고, $\hat{Y}_i$ 는 i번째 입력 데이터에 대한 예측된 유전체 발현 반응 값의 평균값이다.

MAE, MSE 및 CD 점수가 낮으면, 해당 모델이 다른 모델에 비해 더 나은 예측 성능을 달성한다는 것을 나타내고, PC와 R² 점수가 높을수록 해당 모델이 다른 모델에 더 나은 예측 성능을 달성한다는 것을 나타낸다.

도 7은 본 발명의 일 실시예에 따라 4개의 테스트용 데이터 세트를 이용한 유전체 반응 예측 모델을 포함한 기계 학습 기반 베이스 라인 모델의 성능을 나타낸 도면이다.

도 7을 참조하면, 기계 학습 기반 베이스라인 모델은 DNPP(Drug Neighbor Profile Prediction)나 텐서 완성 모델(tensor completion model)인 SVDimpute와 달리, (구 약물, 구 세포주), (신 약물, 구 세포주), (구 약물, 신 세포주), (신 약물, 신 세포주)의 데이터 세트의 후처리된 유전자 발현 반응 정보를 예측할 수 있다.

기계 학습 기반 베이스라인 모델 중에서, TRNet은 모든 평가 지표의 관점에서 모든 종류의 시험 세트에서 최고 성능을 달성하고 있다. TRNet은 MSE 점수가 (구 약물, 구 세포주) ≈ 1051.9점, (신 약물, 구 세포주) ≈ 1181.4점, (구 약물, 신 세포주) ≈ 1218.7점, (신 약물, 신 세포주) ≈ 1306.0점을 각각 달성하고 있음을 알 수 있다.

베이스라인 모델 중에서는 kNN과 RF가 성능이 우수하지만, TRNet의 MSE 점수는 kNN보다 14.88% 높게 나타나고 있다. TRNet에서 얻은 피어슨(Pearson) 상관 계수(PCs)는 0.9에 근접한 값을 가지고 있으므로 약물 투입 후의 유전자 발현 프로파일과 TRNet에 의해 예측된 유전자 발현 반응 정보 사이에 상당한 상관관계가 있음을 나타낸다. TRNet은 입력층에서 세포주와 약물의 양상을 구별하는 심층 신경 네트워크 아키텍처이므로 MLP회귀 분석기의 성능을 능가함을 알 수 있다.

본 발명의 일 실시예에 따른 후보 물질의 발굴 시스템은 TRNet구조로 가지므로 다른 기계 학습 기반 베이스라인 모델과 달리 약물 임베딩 벡터 및 세포주 임베딩 벡터를 산출한다.

TRNet에 의해 생성된 약물 임베딩 벡터를 사용하여 약물 발굴 파이프라인을 시뮬레이션할 수 있다. 임베딩 벡터에는 약물의 전사적 반응과 구조 정보를 포함하고 있다. 비록 2개의 약물이 화학적 구조가 다르더라도 2개의 약

물로 인한 유전자 발현 패턴이 유사한 경우, 2개의 약물은 유사한 임베딩 벡터를 갖게 된다. 따라서, 화합물의 임베딩 벡터가 프로토타입 약물의 임베딩 벡터(알려진 효과를 가진 CMap 화합물)와 유사할 경우, 해당 화합물은 프로토타입 약물과 유사한 전사적 반응을 유도할 수 있으며, 프로토타입 약물과 유사한 효과를 낼 수 있는 후보 물질로 간주할 수 있다.

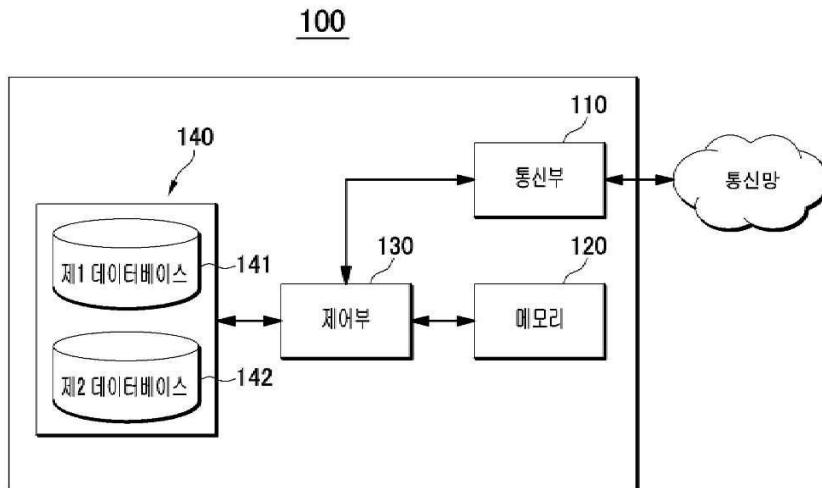
- [0115] ZINC15 화합물과 프로토타입의 약물 임베딩 벡터는 훈련된 TRNet에 의해 획득된다. 여기서, ZINC15 화합물은 제 2 데이터베이스(142)에 의해 수집 및 관리된다. 시제품 약물의 임베딩 벡터와 ZINC 화합물의 임베딩 벡터 사이의 피어슨 상관관계를 계산하고, ZINC15 화합물은 상관 계수에 따라 순위를 설정한다. 정신분열증을 치료하는 약물인 플루페나진(Flupenazine)의 임베딩 벡터와 유사한 임베딩 벡터를 내장하고 있는 상위 등급의 ZINC15 화합물들을 예측할 수 있다.
- [0116] 도 8은 본 발명의 일 실시예에 따라 플루페나진의 임베딩 벡터와 유사한 임베딩 벡터를 내장한 ZINC15 화합물의 상위 10개 후보 물질들을 열거한 도면이다.
- [0117] 도 8을 참조하면, 상위권 후보 물질의 피어슨 상관관계에 기반하여 유사성 점수를 보여준다. 또, 플루페나진의 ECFP 벡터의 자카드(Jaccard) 유사 점수(또는 타니모토 계수 점수)와 후보 물질도 표시된다. 자카드 유사 점수가 높을수록 두 화합물의 구조도 유사하다. 두 화합물 사이의 연관성은 BEST라고 불리는 생물학적 엔티티 검색 도구를 사용하여 측정된다. BEST 도구는 두 화합물이 출판된 문헌의 동일한 추상화에서 함께 언급되는 빈도를 계산하는 것이다. 도 8의 설명란에는 ZINC15 화합물과 플루페나진의 연관성에 대해 설명하고 있다.
- [0118] 문헌 조사에 따르면 10개의 후보 물질 중 7개의 후보 물질이 신경정신과 약물로 사용할 수 있다. 신경정신과 화합물은 플루페나진과의 구조 유사성이 낮음에도 불구하고 플루페나진에게 유사한 전사적 반응을 유도하기 위해 후보 물질로 선정된다. 이와 같이, 학습된 TRNet에 의해 생성된 임베딩 벡터를 포함하는 약물이 전사적 반응 정보에 의해 생성된다는 것을 알 수 있다.
- [0119] 10개의 후보 물질 중에서 플루티젠올, 발프로필벤질라미드, 이미노틸베네에 대해 문헌 조사에서 언급 부분을 찾을 수 없었지만, 해당 화합물들이 신경정신과 약물로 사용될 수 있는지 검토할 가치가 있다.
- [0120] 본 발명의 일 실시예에 따른 후보 물질의 발굴 시스템 및 그 방법은 TR Net을 통해 Cmap 화합물이 저장된 제1 데이터베이스의 (구 약물, 구 세포주)에서 누락된 관측 값을 보완할 수 있고, (신 약물, 구 세포주), (구 약물, 신 세포주) 및 (신 약물, 신 세포주)의 약물에 의한 유전자 발현 반응 정보를 예측할 수 있다.
- [0121] TRNet에서 산출되는 약물 임베딩 벡터와 세포주 임베딩 벡터를 약물 반응 예측, 액션 예측, 약물 조합 시너지 예측, 약물 발굴 등 다양한 약물 개발 과제에 사용될 수 있다. 특히, 세포주 임베딩 벡터는 약물에 대한 반응이 같은 클러스터링 환자나 암 서브타입 분류에 사용될 수 있다.
- [0123] 이상에서 설명한 본 발명의 실시예에 따른 신약 후보 물질의 발굴 방법은, 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터에 의해 실행 가능한 명령어를 포함하는 기록 매체의 형태로도 구현될 수 있다. 이러한 기록 매체는 컴퓨터 판독 가능 매체를 포함하며, 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 판독가능 매체는 컴퓨터 저장 매체를 포함하며, 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.
- [0124] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 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.
- [0125] 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

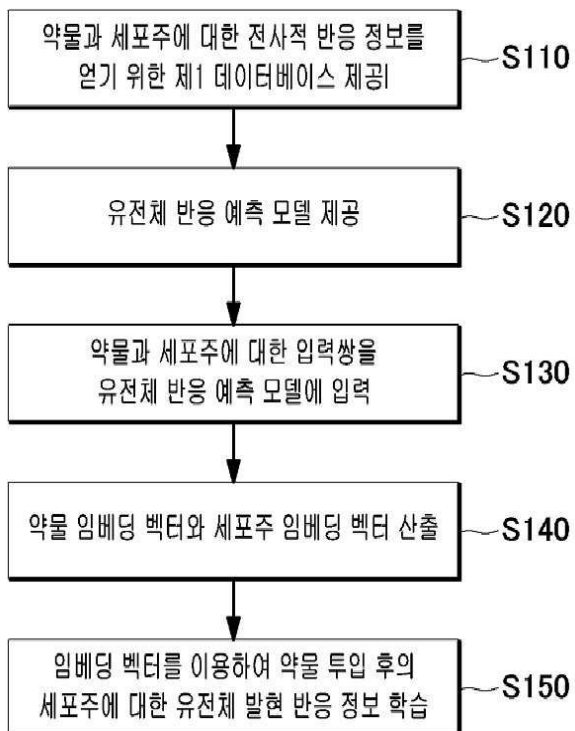
[0126] 100: 신약 후보 물질의 발굴 시스템
 110: 통신 모듈 120: 메모리
 130: 프로세서 140: 데이터베이스
 300: 유전체 반응 예측 모델

도면

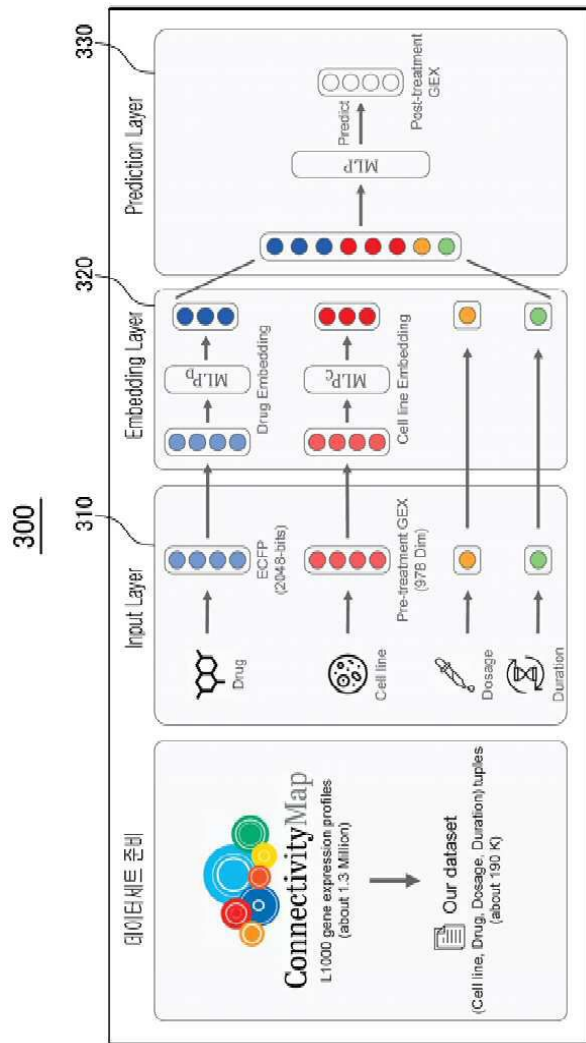
도면1



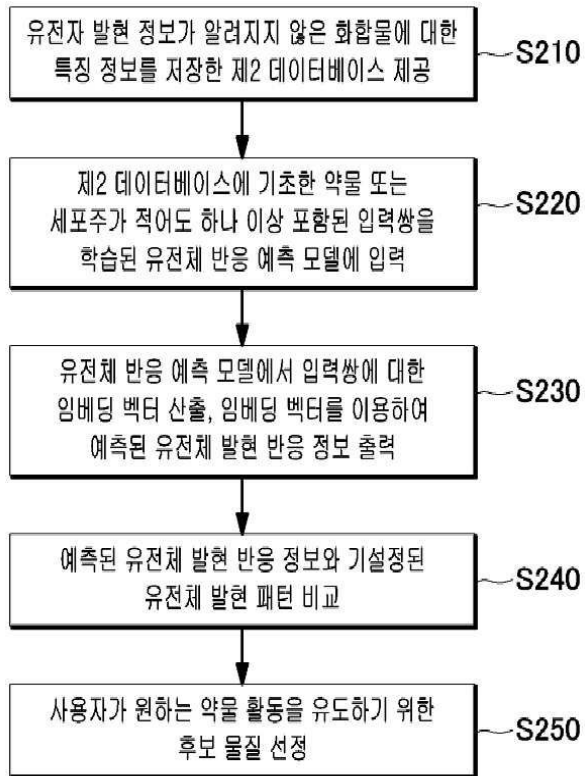
도면2



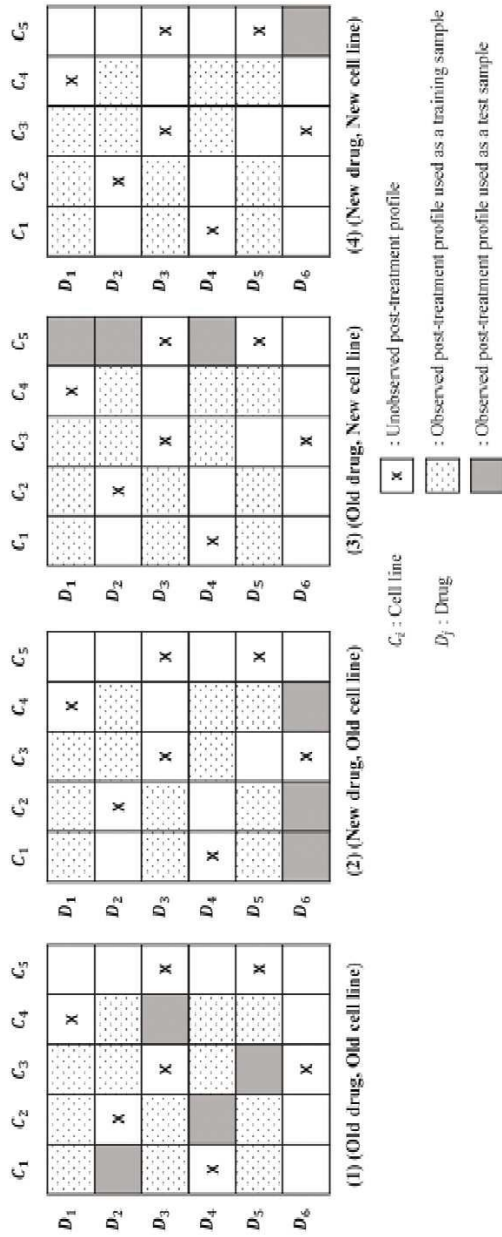
도면3



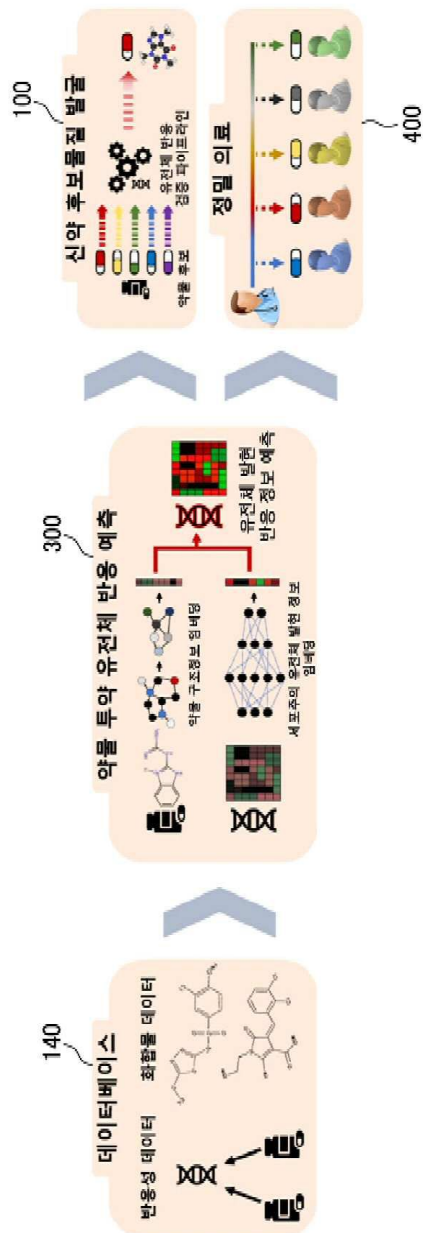
도면4



도면5



도면6



도면7

Test Set	Metric	Models							Improv
		RF	kNN	Ridge	XT	MLP	DNPP	SV/Dimpute	TRNet
(Old drug & Old cell line)	MAE	776.756	752.0402	820.0229	788.4595	792.0580	776.6002	2426.261	670.7712
	MSE	1291.112	1235.8461	1398.9826	1321.8394	1316.4138	1328.8785	10847.2135	1051.9131
	CD	0.0083	0.0079	0.0088	0.0084	0.0084	0.0085	0.0117	0.0069
	PC	0.8964	0.9014	0.8894	0.8941	0.8943	0.8935	0.8594	0.9149
	R ²	0.7296	0.7445	0.7003	0.7216	0.7245	0.735	-0.7719	0.7872
(New drug & Old cell line)	MAE	769.166	744.1804	829.0687	777.3415	802.6100	-	-	719.902
	MSE	1273.1383	1219.41	1418.1413	1291.9866	1356.7430	-	-	1181.3883
	CD	0.0081	0.0078	0.0089	0.0082	0.0086	-	-	0.0076
	PC	0.8983	0.903	0.8881	0.897	0.8921	-	-	0.9061
	R ²	0.7331	0.7501	0.697	0.7256	0.7205	-	-	0.7579
(Old drug & New cell line)	MAE	815.0916	810.96124	837.0687	819.7573	874.9758	-	-	751.7262
	MSE	1374.8737	1358.9191	1419.9407	1391.628	1521.7231	-	-	1218.7394
	CD	0.0087	0.0086	0.0089	0.0088	0.0096	-	-	0.0078
	PC	0.8867	0.8887	0.8841	0.8854	0.8748	-	-	0.9000
	R ²	0.7058	0.7103	0.6922	0.7033	0.6678	-	-	0.7421
(New drug & New cell line)	MAE	806.9333	809.8705	840.7735	811.7987	881.3637	-	-	780.5229
	MSE	1362.2327	1362.1394	1442.9979	1379.901	1550.6442	-	-	1305.9752
	CD	0.0085	0.0085	0.0089	0.0087	0.0097	-	-	0.0082
	PC	0.8884	0.8889	0.8828	0.887	0.8732	-	-	0.8947
	R ²	0.7098	0.7095	0.6868	0.7075	0.6618	-	-	0.7193

도면8

ZINC15 ID	ZINC15 name	Pearson correlation of embedding vectors	Similarity score by ECFP	# of articles	Description
ZINC19366089	Homofenazine	0.986	0.938	0	Homofenazine is an antipsychotic drug and an antixiolytic drug ^a
ZINC22463185	Flutizenol	0.964	0.755	0	-
ZINC19418959	Trifluoperazine	0.953	0.816	139	Trifluoperazine is a antipsychotic agent to treat schizophrenia (Marques and Silva de Lima, 2004)
ZINC21981303	Opipramol	0.933	0.510	0	Opipramol is anxiolytic and antidepressant agent (Jepson and Beaumont, 1973)
ZINC1567421	Mequitazina	0.917	0.322	0	Mequitazina and Fluphenazine are Phenothiazine-derived drugs ^b
ZINC3831041	Mesoridazine	0.917	0.339	15	Mesoridazine is a Phenothiazine-derived drug and used for schizophrenia ^c
ZINC3008621	Valproic	0.911	0.348	5	Valproic is a therapeutic for bipolar disorder ^d
ZINC50	Aptazapine	0.908	0.348	0	Aptazapine has effects for antidepressant ^e
ZINC8714259	Valproylbenzylamide	0.905	0.034	0	-
ZINC1036792	Iminostilbene	0.904	0.171	1	-