



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2016-0095704
(43) 공개일자 2016년08월12일

(51) 국제특허분류(Int. Cl.)

G06F 19/00 (2011.01)

(52) CPC특허분류

G06F 19/705 (2013.01)

G06F 19/706 (2013.01)

(21) 출원번호 10-2015-0016938

(22) 출원일자 2015년02월03일

심사청구일자 2015년02월03일

(71) 출원인

이화여자대학교 산학협력단

서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)

(72) 발명자

김완규

서울특별시 송파구 올림픽로 99 109동 1801호 (잠실동, 잠실엘스아파트)

이해승

서울특별시 서대문구 신촌로 109 915호 (창천동, 르메이에르5차)

(74) 대리인

손민

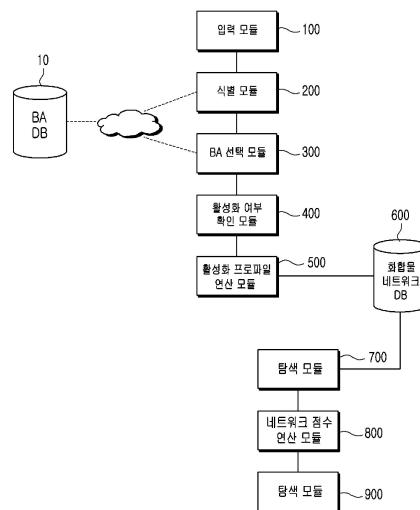
전체 청구항 수 : 총 15 항

(54) 발명의 명칭 화합물 네트워크 데이터베이스 구축 방법과 약물 가상 탐색 방법 및 이를 위한 시스템

(57) 요약

본 발명은, 사용자가 입력한 화합물과 유사한 화합물을 높은 정확도로 확인할 수 있는 데이터베이스를 구축하고, 이를 통하여 약물 가상 탐색을 신속하고 정확하게 수행할 수 있는 방법과 시스템에 관한 것이다.

대표도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호 1711013658

부처명 미래창조과학부

연구관리전문기관 한국연구재단

연구사업명 기초연구사업(학술진흥)(2014)-중견연구자지원사업-핵심연구지원사업(개인)

연구과제명 전사체 정보 기반의 항암제 반응성에 따른 아교모세포종 subtype 분류 및 조합 약물 예측
을 위한 생명정보 분석 도구 개발

기 여 율 1/1

주관기관 이화여자대학교 산학협력단

연구기간 2014.05.01 ~ 2017.04.30

명세서

청구범위

청구항 1

- (a) 식별 모듈이, 입력 모듈을 통해 입력된 한 쌍의 화합물의 식별자(ID1, ID2)를 각각 식별하는 단계;
- (b) 바이오어세이 선택 모듈이, 바이오어세이 데이터베이스에서 상기 식별자(ID1, ID2)에 해당하는 화합물이 포함된 바이오어세이들을 각각 선택하는 단계;
- (c) 활성화 여부 확인 모듈이, 기 설정된 방법을 이용하여, 상기 선택된 바이오어세이들에서 이에 대응하는 화합물에 대한 활성화 여부를 확인하는 단계;
- (d) 활성화 프로파일 연산 모듈이, 상기 (c) 단계에서 확인된 활성화 여부를 이용하여 상기 입력된 한 쌍의 화합물의 유사도 점수(similarity score)를 연산하는 단계;
- (e) 상기 활성화 프로파일 연산 모듈이, 상기 연산된 유사도 점수가 기 설정된 기준값(cut-off value) 이상인 경우, 상기 한 쌍의 화합물의 유사한 것으로 판단하는 단계; 및
- (f) 상기 (e) 단계에서 판단된 유사 여부 및 상기 (d) 단계에서 연산된 유사도 점수가 화합물 네트워크 데이터베이스에 저장되는 단계를 포함하는, 화합물 네트워크 데이터베이스 구축 방법.

청구항 2

- 제 1 항에 있어서,
- 상기 (f) 단계 이후,
- (g) 다른 한 쌍의 화합물이 추가로 선택된 경우, 상기 선택된 다른 한 쌍의 화합물에 대하여 상기 (a) 내지 상기 (f) 단계가 반복되는 단계를 더 포함하는, 화합물 네트워크 데이터베이스 구축 방법.

청구항 3

- 제 2 항에 있어서,
- 상기 (g) 단계 이후, 반복된 상기 (a) 내지 상기 (f) 단계의 결과에 따라, 어느 하나의 화합물마다 각각 유사한 것으로 판단된 다른 하나 이상의 화합물과 비유사한 것으로 판단된 다른 하나 이상의 화합물이 상기 화합물 네트워크 데이터베이스에 저장되는, 화합물 네트워크 데이터베이스 구축 방법.

청구항 4

- 제 1 항에 있어서,
- 상기 (d) 단계는, 코사인 유사도(cosine similarity), 유클리디안 거리(Euclidean distance), 스칼라의 곱(inner product), 피어슨 계수(Pearson coefficient), 스피어만 계수(spearman coefficient), 상호 정보(mutual information), 자카드 유사도(Jaccard similarity), 및 타니모토 계수(Tanimoto coefficient) 중 어느 하나의 방법을 이용하여 수행되는,

화합물 네트워크 데이터베이스 구축 방법.

청구항 5

제 1 항 내지 제 4 항 중 어느 한 항에 따른 화합물 네트워크 데이터베이스 구축 방법을 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 매체.

청구항 6

제 1 항 내지 제 4 항 중 어느 한 항에 따른 방법에 따라 구축된 상기 화합물 네트워크 데이터베이스를 이용한 약물 가상 탐색 방법으로서,

- (h) 상기 식별 모듈이, 상기 입력 모듈을 통해 입력된 화합물의 식별자(ID)를 식별하는 단계; 및
 - (i) 탐색 모듈이, 상기 화합물 네트워크 데이터베이스에서 상기 식별자(ID)에 해당하는 화합물과 유사한 것으로 저장된 유사 화합물들을 탐색하는 단계를 포함하는,
- 약물 가상 탐색 방법.

청구항 7

제 6 항에 따른 약물 가상 탐색 방법을 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 매체.

청구항 8

제 1 항 내지 제 4 항 중 어느 한 항에 따른 방법에 따라 구축된 상기 화합물 네트워크 데이터베이스를 이용한 약물 가상 탐색 방법으로서,

- (h') 상기 식별 모듈이, 상기 입력 모듈을 통해 입력된 다수의 화합물의 식별자들(ID1~IDn)을 식별하는 단계; 및
 - (i') 탐색 모듈이, 상기 화합물 네트워크 데이터베이스에서 상기 식별자들(ID1~IDn)에 해당하는 화합물들과 유사한 것으로 저장된 유사 화합물들을 모두 탐색하는 단계를 포함하는,
- 약물 가상 탐색 방법.

청구항 9

제 8 항에 있어서,

상기 (i') 단계 이후,

- (j) 네트워크 점수 연산 모듈이, 상기 유사 화합물들을 기준으로 상기 식별자들(ID1~IDn)에 해당하는 화합물들 각각과의 유사도 점수를 통합함으로써, 상기 유사 화합물들에 대한 네트워크 점수(network score)를 각각 연산하는 단계; 및
 - (k) 출력 모듈이, 상기 유사 화합물들을 상기 연산된 네트워크 점수가 높은 순서에 따라 정렬하여, 정렬된 순서에 따라 상기 유사 화합물이 상기 (h') 단계에서 입력된 다수의 화합물과 더 유사한 화합물인 것으로 출력하는 단계를 더 포함하는,
- 약물 가상 탐색 방법.

청구항 10

제 9 항에 따른 약물 가상 탐색 방법을 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 매체.

청구항 11

화합물이 입력되는 입력 모듈;

상기 입력 모듈을 통해 입력된 화합물의 식별자를 식별하는 식별 모듈;

바이오어세이 데이터베이스에서 식별자에 해당하는 화합물이 포함된 바이오어세이들을 선택하는 바이오어세이 선택 모듈;

기 설정된 방법을 이용하여, 상기 선택된 바이오어세이들이 해당 화합물에 대하여 활성화인지 여부를 확인하는 활성화 여부 확인 모듈;

한 쌍의 화합물이 상기 입력 모듈을 통하여 입력된 경우, 상기 한 쌍의 화합물에 대한 활성화 여부를 이용하여 상기 입력된 한 쌍의 화합물의 유사도 점수(similarity score)를 연산하고, 상기 연산된 유사도 점수가 기 설정된 기준값 이상인 경우 상기 입력된 한 쌍의 화합물이 유사한 것으로 판단하는, 활성화 프로파일 연산 모듈; 및

상기 활성화 프로파일 연산 모듈에서 연산된 유사도 점수와 판단된 유사 여부가 저장되는 화합물 네트워크 데이터베이스를 포함하는,

화합물 네트워크 데이터베이스 구축 시스템.

청구항 12

제 11 항에 있어서,

상기 활성화 프로파일 연산 모듈은, 코사인 유사도(cosine similarity), 유클리디안 거리(Euclidean distance), 스칼라의 곱(inner product), 피어슨 계수(Pearson coefficient), 스피어만 계수(spearman coefficient), 상호 정보(mutual information), 자카드 유사도(Jaccard similarity), 및 타니모토 계수(Tanimoto coefficient) 중 어느 하나의 방법을 이용하여 유사도 점수를 연산하는,

화합물 네트워크 데이터베이스 구축 시스템.

청구항 13

제 11 항 또는 제 12 항에 따른 화합물 네트워크 데이터베이스 구축 시스템에서 구축된 상기 화합물 네트워크 데이터베이스; 및

어느 하나의 화합물이 상기 입력 모듈을 통하여 입력된 경우, 상기 화합물 네트워크 데이터베이스에서 상기 입력된 화합물과 유사한 것으로 저장된 유사 화합물들을 탐색하는 탐색 모듈을 포함하는,

약물 가상 탐색 시스템.

청구항 14

제 11 항 또는 제 12 항에 따른 화합물 네트워크 데이터베이스 구축 시스템에서 구축된 상기 화합물 네트워크 데이터베이스; 및

다수의 화합물이 상기 입력 모듈을 통하여 입력된 경우, 상기 화합물 네트워크 데이터베이스에서 상기 입력된 다수의 화합물과 유사한 것으로 저장된 유사 화합물들을 탐색하는 탐색 모듈을 포함하는,

약물 가상 탐색 시스템.

청구항 15

제 14 항에 있어서,

상기 유사 화합물들을 기준으로 상기 식별자들(ID1~IDn)에 해당하는 화합물들 각각과의 유사도 점수를 통합함으로써, 상기 유사 화합물들에 대한 네트워크 점수를 각각 연산하는, 네트워크 점수 연산 모듈; 및

상기 유사 화합물들을 상기 연산된 네트워크 점수가 높은 순서에 따라 정렬하여, 정렬된 순서에 따라 상기 유사 화합물이 상기 입력 모듈에 입력된 다수의 화합물과 더 유사한 화합물인 것으로 출력하는 출력 모듈을 더 포함하는,

약물 가상 탐색 시스템.

발명의 설명

기술 분야

[0001] 본 발명의 기술 분야는 바이오인포매틱스이다. 구체적으로, 본 발명은 적은 시간 및 노력으로도 방대한 데이터베이스 내에서 입력한 화합물과 유사 화합물을 높은 정확도로 확인할 수 있는 데이터베이스를 구축하고, 이를 이용하여 약물 가상 탐색을 하는 방법과 시스템에 관한 것이다.

배경 기술

[0002] 최근 신약개발과정에 있어서, 시간과 비용을 줄일 수 있는 수단으로서 컴퓨터를 이용한 가상 탐색 기술(Virtual Screening)이 핵심 분야 중 하나로 자리 잡고 있다. 그 배경에는 계산기의 비약적인 성능 향상에 가세해 고속 다중 탐색 기술(High Throughput Screening)이나 조합 합성 기술(Combinatorial Synthesis)에 의해 축적되는 구조-활성 관계(Structure-Activity Relationship)의 정보량이 비약적으로 증대하고 있는 점, 게놈 연구의 진전에 의해 타겟 단백질의 구조 정보가 비약적으로 증대하고 있는 점 등을 들 수 있다.

[0003] 이와 같은 가상 스크리닝 기술은 전통적으로, 타겟 단백질에 활성이 있는 것으로 알려진 화합물들 간의 구조적 유사성, 즉 이미 알려진 구조-활성 관계 정보에 근거하는 리간드 기반 가상 탐색 방법(Ligand-Based Virtual Screening)과 타겟 단백질의 입체 구조 정보를 이용하여 단백질-리간드 도킹(docking)과 같이 구조 기반 가상 탐색 방법(Structure-Base Virtual Screening)이 있다.

[0004] 구조 기반 탐색 방법은, 복수의 약물이 타겟 단백질의 활성 부위 부근에 결합하는 경우, 어느 약물도 단백질과 상보적인 관계에 있는 것과 동시에, 각각의 결합 과정에 있어서의 자유에너지 변화량이 약리 활성의 강약을 나타낸다는 개념에 근거한다. 타겟 단백질과 리간드와의 결합 상태와 그 약리 활성치를 컴퓨터 위에서 추정하며, 구조-활성 관계의 정보를 필요로 하지 않음에도 불구하고, 정확도 높은 활성치 예측을 기대할 수 있는 이점을 가지고 있다. 그러나, 리간드(true ligand)를 비리간드(non-ligand)와 분별하는 것은 가능하나 정량적인 순서를 매기는 것은 거의 불가능하며, 대부분의 가상탐색/도킹 프로그램은 단백질의 유연성을 고려하지 못하는 한계가 있다. 또한, 수용체 구조(결합 모델)가 필수이며, 예측의 정도(accuracy)는 구조의 정확성에 의존한다. 뿐만 아니라, 구조-활성 관계 정보가 축적되어도 예측 정확도의 향상으로 연결되지 않는다는 한계가 있다.

[0005] 한편, 공통의 부위에 결합하고 있는 약물 간에는 그 물리화학적 파라미터에 있어서 상동성을 볼 수 있다는 점에 착안한 리간드 기반 가상 탐색 방법은 구조 기반 가상 탐색 방법과는 달리 수용체 구조(결합 모델)를 필요로 하지 않는 이점이 있으나, 약리 활성의 사전 정보를 필요로 하고, 사전 정보의 질과 양에 예측 정확도가 의존하고 있기 때문에, 기지 정보 이상의 예측은 불가능하거나, 활성치 예측의 정확도가 낮다는 한계가 여전히 존재한다.

[0006] 본 발명자들은 상기와 같은 기존 가상 탐색 기술의 한계를 극복하기 위해 연구를 거듭한 결과, 기존의 방법과는 달리, 사용자가 입력한 화합물과 유사한 화합물을 유사한 순서대로 출력할 수 있는 화합물 네트워크 데이터베이스를 구축하고, 이를 통하여 신속하고 정확하게 약물 가상 탐색을 이룰 수 있는 방법을 제안하기에 이르렀다.

[0007] (특허문헌 1) US6,421,612

[0008] (특허문헌 2) US6,994,473

[0009] (특허문헌 3) US7,416,524

발명의 내용

해결하려는 과제

[0010] 본 발명은 상기와 같은 과제를 해결하기 위하여 안출된 것이다.

[0011] 구체적으로, 방대한 양의 데이터베이스를 사용하기에 발생하는 시간, 노력의 낭비의 문제를 해결함과 동시에 정확성 있는 가상 유사 화합물을 출력할 수 있는 방법 및 시스템을 제안하고자 한다.

과제의 해결 수단

[0012] 상기와 같은 과제를 해결하기 위하여, 본 발명의 일 실시예는, (a) 식별 모듈이, 입력 모듈을 통해 입력된 한 쌍의 화합물의 식별자(ID1, ID2)를 각각 식별하는 단계; (b) 바이오어세이 선택 모듈이, 바이오어세이 데이터베이스에서 상기 식별자(ID1, ID2)에 해당하는 화합물이 포함된 바이오어세이들을 각각 선택하는 단계; (c) 활성화 여부 확인 모듈이, 기 설정된 방법을 이용하여, 상기 선택된 바이오어세이들에서 이에 대응하는 화합물에 대한 활성화 여부를 확인하는 단계; (d) 활성화 프로파일 연산 모듈이, 상기 (c) 단계에서 확인된 활성화 여부를 이용하여 상기 입력된 한 쌍의 화합물의 유사도 점수(similarity score)를 연산하는 단계; (e) 상기 활성화 프로파일 연산 모듈이, 상기 연산된 유사도 점수가 기 설정된 기준값(cut-off value) 이상인 경우, 상기 한 쌍의 화합물의 유사한 것으로 판단하는 단계; 및 (f) 상기 (e) 단계에서 판단된 유사 여부 및 상기 (d) 단계에서 연산된 유사도 점수가 화합물 네트워크 데이터베이스에 저장되는 단계를 포함하는, 화합물 네트워크 데이터베이스 구축 방법을 제공한다.

[0013] 또한, 상기 (f) 단계 이후, (g) 다른 한 쌍의 화합물이 추가로 선택된 경우, 상기 선택된 다른 한 쌍의 화합물에 대하여 상기 (a) 내지 상기 (f) 단계가 반복되는 단계를 더 포함하는 것이 바람직하다.

[0014] 또한, 상기 (g) 단계 이후, 반복된 상기 (a) 내지 상기 (f) 단계의 결과에 따라, 어느 하나의 화합물마다 각각 유사한 것으로 판단된 다른 하나 이상의 화합물과 비유사한 것으로 판단된 다른 하나 이상의 화합물이 상기 화합물 네트워크 데이터베이스에 저장되는 것이 바람직하다.

[0015] 또한, 상기 (d) 단계는, 유사도 점수를 연산하는 어떠한 방법을 사용하여도 무방하나, 예를 들어 코사인 유사도(cosine similarity), 유클리디안 거리(Euclidean distance), 스칼라의 곱(inner product), 피어슨 계수(Pearson coefficient), 스피어만 계수(spearman coefficient), 상호 정보(mutual information), 자카드 유사도(Jaccard similarity), 및 타니모토 계수(Tanimoto coefficient) 등의 방법을 사용할 수 있다.

[0016] 상기와 같은 과제를 해결하기 위하여, 본 발명의 다른 실시예는, 전술한 방법에 따라 화합물 네트워크 데이터베이스를 구축한 후, (h) 상기 식별 모듈이, 상기 입력 모듈을 통해 입력된 화합물의 식별자(ID)를 식별하는 단계; 및 (i) 탐색 모듈이, 상기 화합물 네트워크 데이터베이스에서 상기 식별자(ID)에 해당하는 화합물과 유사한 것으로 저장된 유사 화합물들을 탐색하는 단계를 포함하는, 약물 가상 탐색 방법을 제공한다.

- [0017] 상기와 같은 과제를 해결하기 위하여, 본 발명의 또 다른 실시예는, 전술한 방법에 따라 화합물 네트워크 데이터베이스를 구축한 후, (h') 상기 식별 모듈이, 상기 입력 모듈을 통해 입력된 다수의 화합물의 식별자들(ID1~IDn)을 식별하는 단계; 및 (i') 탐색 모듈이, 상기 화합물 네트워크 데이터베이스에서 상기 식별자들(ID1~IDn)에 해당하는 화합물들과 유사한 것으로 저장된 유사 화합물들을 모두 탐색하는 단계를 포함하는, 약물 가상 탐색 방법을 제공한다.
- [0018] 또한, 상기 (i') 단계 이후, (j) 네트워크 점수 연산 모듈이, 상기 유사 화합물들을 기준으로 상기 식별자들(ID1~IDn)에 해당하는 화합물들 각각과의 유사도 점수를 통합함으로써, 상기 유사 화합물들에 대한 네트워크 점수(network score)를 각각 연산하는 단계; (k) 출력 모듈이, 상기 유사 화합물들을 상기 연산된 네트워크 점수가 높은 순서에 따라 정렬하여, 정렬된 순서에 따라 상기 유사 화합물이 상기 (h') 단계에서 입력된 다수의 화합물과 더 유사한 화합물인 것으로 출력하는 단계를 더 포함하는 것이 바람직하다.
- [0019] 상기와 같은 과제를 해결하기 위하여, 본 발명의 또 다른 실시예는, 화합물이 입력되는 입력 모듈; 상기 입력 모듈을 통해 입력된 화합물의 식별자를 식별하는 식별 모듈; 바이오어세이 데이터베이스에서 식별자에 해당하는 화합물이 포함된 바이오어세이드를 선택하는 바이오어세이 선택 모듈; 기 설정된 방법을 이용하여, 상기 선택된 바이오어세이드들이 해당 화합물에 대하여 활성화인지 여부를 확인하는 활성화 여부 확인 모듈; 한 쌍의 화합물이 상기 입력 모듈을 통하여 입력된 경우, 상기 한 쌍의 화합물에 대한 활성화 여부를 이용하여 상기 입력된 한 쌍의 화합물의 유사도 점수(similarity score)를 연산하고, 상기 연산된 유사도 점수가 기 설정된 기준값 이상인 경우 상기 입력된 한 쌍의 화합물이 유사한 것으로 판단하는, 활성화 프로파일 연산 모듈; 및 상기 활성화 프로파일 연산 모듈에서 연산된 유사도 점수와 판단된 유사 여부가 저장되는 화합물 네트워크 데이터베이스를 포함하는, 화합물 네트워크 데이터베이스 구축 시스템을 제공한다.
- [0020] 또한, 상기 활성화 프로파일 연산 모듈은, 유사도 점수를 연산하는 어떠한 방법을 사용하는 모듈이어도 무방하나, 예를 들어 코사인 유사도(cosine similarity), 유클리디안 거리(Euclidean distance), 스칼라의 곱(inner product), 피어슨 계수(Pearson coefficient), 스피어만 계수(spearman coefficient), 상호 정보(mutual information), 자카드 유사도(Jaccard similarity), 및 타니모토 계수(Tanimoto coefficient) 등의 방법을 사용할 수 있다.
- [0021] 상기와 같은 과제를 해결하기 위하여, 본 발명의 또 다른 실시예는, 전술한 화합물 네트워크 데이터베이스; 및 어느 하나의 화합물이 상기 입력 모듈을 통하여 입력된 경우, 상기 화합물 네트워크 데이터베이스에서 상기 입력된 화합물과 유사한 것으로 저장된 유사 화합물들을 탐색하는 탐색 모듈을 포함하는, 약물 가상 탐색 시스템을 제공한다.
- [0022] 상기와 같은 과제를 해결하기 위하여, 본 발명의 또 다른 실시예는, 전술한 화합물 네트워크 데이터베이스; 및 다수의 화합물이 상기 입력 모듈을 통하여 입력된 경우, 상기 화합물 네트워크 데이터베이스에서 상기 입력된 다수의 화합물과 유사한 것으로 저장된 유사 화합물들을 탐색하는 탐색 모듈을 포함하는, 약물 가상 탐색 시스템을 제공한다.
- [0023] 또한, 상기 유사 화합물들을 기준으로 상기 식별자들(ID1~IDn)에 해당하는 화합물들 각각과의 유사도 점수를 통합함으로써, 상기 유사 화합물들에 대한 통합된 네트워크 점수를 각각 연산하는, 네트워크 점수 연산 모듈; 및 상기 유사 화합물들을 상기 연산된 네트워크 점수가 높은 순서에 따라 정렬하여, 정렬된 순서에 따라 상기 유사 화합물이 상기 입력 모듈에 입력된 다수의 화합물과 더 유사한 화합물인 것으로 출력하는 출력 모듈을 더 포함하는 것이 바람직하다.

발명의 효과

- [0024] 본 발명에 의하여, 소수의 입력 세트에 속한 화합물들과 그 생리적, 화학적 활성이 유사할 것으로 예측되는 화합물을 예측함으로써 생물학적 검정(bioassay) 실험의 성공 수율(yield rate)을 향상시키고, 이를 통해 집중 탐색용 화합물 라이브러리를 설계하는데 활용될 수 있다.

[0025] 또한, 다수의 다양한 타겟 단백질에 작용하는 것으로 알려진 화합물 세트들을 입력 화합물 세트로 하여 후보 화합물들의 네트워크 점수를 비교함으로써, 그 타겟이 알려져 있지 않은 특정 화합물에 대한 타겟을 추정하는 데에도 적용된다.

[0026] 결과적으로, 본 발명은 표현형 기반 약물 탐색(phenotypic drug screening)을 통해 찾아낸 화합물들과 같이 그 타겟의 수와 종류가 정확하게 알려지지 않은 경우에도 적용이 가능하며, 그 화합물 세트에 유의한 실험의 방법 및 내용을 통해 해당 표현형과 관련된 미지의 타겟을 찾아내는 데에도 적용할 수 있다.

도면의 간단한 설명

[0027] 도 1은 본 발명에 따른 시스템을 설명하는 개념도이다.

도 2는 본 발명에 따른 화합물 네트워크 데이터베이스 구축 방법을 설명하는 순서도이다.

도 3은 본 발명에 따른 약물 가상 탐색 방법을 설명하는 순서도이다.

도 4는 본 발명의 설명을 참조하기 위한 개략도이다.

도 5 및 도 6은 본 발명을 검증하기 위한 실험의 데이터 및 결과인 ROC를 각각 도시한다.

발명을 실시하기 위한 구체적인 내용

[0028] 1. 용어의 정의

[0029] 이하에서 "화합물(compound)"은 공지된 어떠한 화합물도 가능하다. 각각의 화합물에는 식별 및 다른 화합물과의 구분을 위한 식별자(ID; identifier)가 할당된다. 식별자를 표기하는 방법은 이미 알려진 표준에 의한다.

[0030] 이하에서 "바이오어세이(bioassay)"는, 이미 실행되어 알려진 검정실험의 결과에 대한 데이터를 의미한다. 바이오어세이에는 다수의 화합물이 할당되며, 반대로 특정 화합물에 상응하는 바이오어세이 역시 다수일 수 있다.

[0031] 이하에서 "가상 탐색(Virtual Screening)"은, 전통적으로 약물이나 타겟의 구조 정보 등을 이용하여 특정 단백질 타겟에 활성이 있을 것으로 예상되는 화합물을 예측하는 방법을 지칭한다.

[0032] 이하에서 "활성도(activity)"는 특정 조건(예를 들어, 표적 내지 특정 단백질)에 대하여 특정 화합물이 반응을 나타내는 정도를 의미한다. 활성도는 다양한 지표로 표현되는데 본 발명에서는 정량화되어 입력됨을 가정하며, 이러한 방식 역시 종래의 일반적 방식인바 상세한 설명은 생략한다. 정량화된 활성 정도에 따라 특정 화합물은 특정 조건에서 활성이거나 비활성일 수 있어서, 활성화 여부에 대한 객관적이고 보편적 검증이 가능하다.

[0033] 이하에서 "활성 프로파일 (activity profile)"은 특정 화합물을 대상으로 실험된 모든 생물학적 검정(bioassay) 데이터를 벡터 형식으로 표현한 것이다. 따라서 벡터의 크기는 특정 화합물이 검정된 실험의 숫자와 동일하다. 프로파일을 구성하는 데이터는 실험 검정 결과 값 또는 이를 정규화(Normalization)한 값을 그대로 사용하거나, 활성 여부에 따라 1과 0으로 전환한 값을 사용한다.

[0034] 이하에서 "활성 프로파일 유사도(activity profile similarity)"는 화합물의 활성 프로파일을 이용하여 서로 다른 두 화합물의 닮은 정도를 나타낸다. 두 화합물의 활성 상태가 비슷할수록 유사도는 높아진다.

[0035] 이하에서 "화합물-화합물 네트워크(Compound-Compound network)" 또는 "화합물 네트워크"는 화합물과 화합물간의 관계를 활성프로파일 유사도 기반으로 연결한 상호 네트워크를 의미한다.

[0036] 이하에서, "ROC 곡선(Receiver-Operating Characteristic Curve)" 특정 예측 방법의 효율성을 확인하기 위하여 적중 확률, 즉, 민감도(sensitivity)를 Y축으로 하고 오경보 확률, 즉, 1-특이도(1-specificity)를 X축으로 한 곡선을 의미하며, 또한 여기에서 AUC(Area Under Curve) 값은 곡선하 면적값을 의미한다.

[0037] 이하에서 "시스템(system)"은 방법의 반대 개념인 물건을 의미하는 것으로 이해되어야 한다.

[0038] 이하에서 "모듈(module)"은 정보 처리를 위한 연산 수단의 단위를 의미하는 것으로서, 각각의 모듈이 반드시 물리적으로 구분될 필요는 없으며, 하나의 모듈은 하나의 수단에서만 이루어질 수도 있고 다수의 모듈이 하나의 수단에서 이루어질 수도 있다. 예를 들어, 본 발명에 따른 시스템이 정보 처리 단말기(즉, 컴퓨터)에서 이루어지는 경우, 하나의 단말기에서 모든 모듈의 연산 작용이 수행될 수도 있으며, 각각의 모듈이 별도의 단말기에서

연산 작용을 수행할 수도 있다.

[0039] 2. 시스템의 설명

[0040] 도 1을 참조하여 본 발명에 따른 시스템을 먼저 설명한다.

[0041] 본 발명에 따른 시스템은 입력 모듈(100), 식별 모듈(200), 바이오어세이 선택 모듈(300), 활성화 여부 확인 모듈(400), 활성화 프로파일 연산 모듈(500), 탐색 모듈(700), 네트워크 점수 연산 모듈(800) 및 출력 모듈(900)을 포함한다.

[0042] 이 중에서 입력 모듈(100), 식별 모듈(200), 바이오어세이 선택 모듈(300), 활성화 여부 확인 모듈(400) 및 활성화 프로파일 연산 모듈(500)은 본 발명 특유의 화합물 네트워크 데이터베이스(600)를 구축하기 위한 시스템의 구성이다.

[0043] 또한, 구축된 화합물 네트워크 데이터베이스(600)를 이용하여 유사 화합물을 탐색하고 출력하기 위하여, 탐색 모듈(700), 네트워크 점수 연산 모듈(800) 및 출력 모듈(900)이 추가된다.

[0044] 또한, 본 발명에 따른 시스템은 웹(web) 등으로 연결되어 별도로 구축된 바이오어세이 데이터베이스(10)에 접속하여 정보를 업로드 및 다운로드할 수 있다.

[0045] 바이오어세이 데이터베이스(10)는, 바이오어세이 데이터가 포함되어 있는 어떠한 데이터베이스도 무방하다. 다만, 바이오어세이 데이터는 전술한 바와 같이 다수의 화합물이 포함되어 있어야 하며, 후술할 바와 같이 식별 모듈(200)에서 식별되는 식별자를 이용하여 화합물이 자동으로 검색될 수 있는 형식(format)으로 저장된 데이터이어야 한다.

[0046] 바이오어세이 데이터베이스(10)의 일례는, 미국보건연구원에서 제공하는 바이오어세이 데이터베이스일 수 있으나(<https://pubchem.ncbi.nlm.nih.gov>), 이에 제한이 없음은 물론이다.

[0047] 입력 모듈(100)은 사용자가 화합물을 입력할 수 있는 어떠한 수단이어도 무방하다. 키보드, 마우스, 터치패드 등이 일례일 수 있다.

[0048] 식별 모듈(200)은 사용자가 입력한 화합물에서 미리 결정된 식별자를 식별한다. 이를 위하여, 바이오어세이 데이터베이스(10)에는 화합물마다 식별자가 매핑(mapping)되어 저장되어 있는 것이 바람직하다. 전술한 일례인, 미국보건연구원에서 제공하는 바이오어세이 데이터베이스에는 이와 같은 식별자가 모두 저장되어 있다. 예를 들어, 타겟 단백질이 "Ubiquitin carboxyl-terminal hydrolase 1"인 경우 그 식별자는 "094782"로 저장되어 있다. 식별 모듈(200)은 웹을 통하여 바이오어세이 데이터베이스(10)에 접속함으로써, 입력 모듈(100)을 통하여 입력된 화합물의 식별자를 자동으로 식별한다.

[0049] 바이오어세이 선택 모듈(300)은 특정 화합물이 포함된 바이오어세이를, 바이오어세이 데이터베이스(10)로부터 선택한다.

[0050] 전술한 바와 같이, 하나의 바이오어세이에 다수의 화합물이 포함되고, 마찬가지로 특정 화합물이 포함된 바이오어세이는 다수 개 존재하는데, 입력 모듈(100)에서 입력되고 식별 모듈(200)에서 식별자가 식별된 특정 화합물이 포함된 다수의 바이오어세이를 바이오어세이 데이터베이스(10)로부터 선택하는 기능을 한다.

[0051] 예를 들어, 도 4의 (a)와 같이 특정 화합물인 "C"에 대하여 8개의 바이오어세이가 선택될 수 있다.

[0052] 활성화 여부 확인 모듈(400)은 다수의 바이오어세이들에서 해당 화합물이 특정 조건에서의 활성화 여부를 확인한다.

[0053] 전술한 바와 같이, 활성화 여부는 종래 널리 알려진 일반적인 방법을 사용하여 정량화할 수 있는데, 활성화 여부 확인 모듈(400)은 미리 설정된 기준을 이용하여 활성(active)과 비활성(inactive)으로 구분한다.

- [0054] 예를 들어, 도 4의 (a)와 같이 특정 화합물인 "C"에 대하여 8개의 바이오어세이가 선택되었으나, 4개의 화합물만 활성이고, 다른 4개의 화합물은 비활성으로 확인되었음을 알 수 있다.
- [0055] 활성화 프로파일 연산 모듈(500)은 한 쌍의 화합물이 입력되고 식별되고 바이오어세이들이 다수 선택되고 각각의 활성화 여부가 확인된 경우, 이들을 이용하여 입력된 한 쌍의 화합물의 유사도 점수(similarity score)를 연산하고, 상기 연산된 유사도 점수가 기 설정된 기준값(cut-off value) 이상인 경우 상기 입력된 한 쌍의 화합물이 유사한 것으로 판단하는 기능을 한다.
- [0056] 활성화 프로파일 연산 모듈(500)은 한 쌍의 화합물이 입력되면 각각에 대하여 선택된 바이오어세이들과 활성화 여부를 비교함으로써 유사도 점수를 연산할 수 있다.
- [0057] 유사도 점수를 연산하는 방법은 종래의 어떠한 통계적 방법을 사용하여도 무방하나, 예를 들어 코사인 유사도(cosine similarity), 유클리디안 거리(Euclidean distance), 스칼라의 곱(inner product), 피어슨 계수(Pearson coefficient), 스피어만 계수(spearman coefficient), 상호 정보(mutual information), 자카드 유사도(Jaccard similarity), 타니모토 계수(Tanimoto coefficient) 등 다양한 통계적 방법을 적용할 수 있다. 위의 방법들은 예시에 불과하며, 벡터 사이의 유사도를 연산하기 위한 어떠한 방법을 사용하여도 무방함을 다시 한 번 강조한다.
- [0058] 예를 들어, 도 4의 (b)와 같이 한 쌍의 화합물인 "A"와 "B"에 대하여 각각 바이오어세이들이 선택되었는데, 9개의 일반적 어세이들 중 5개가 활성화되었음을 확인할 수 있으며, 여기에서 자카드 연산법에 의하여 유사도 점수를 연산하면,
- [0059] $J(A, B) = 5/7$
- [0060] 과 같은 결과를 얻을 수 있다. 얻어진 결과값이 유사도 점수가 된다.
- [0061] 또한, 유사도 점수에 미리 설정된 기준값을 적용하면 유사 여부를 연산할 수 있다. 예를 들어, 도 4의 (b)와 같은 결과에서, 기준값이 0.5인 경우 화합물 "A"와 "B"는 유사한 것으로 연산된다.
- [0062] 활성화 프로파일 연산 모듈(500)이 다수의 화합물 쌍에 대하여 반복 수행한다면 특정 한 쌍의 화합물마다 유사 여부가 결정되며, 결정된 유사 여부는 화합물 네트워크 데이터베이스(600)에 저장될 수 있다.
- [0063] 예를 들어, 도 4의 (C)에서 다수의 원은 서로 다른 화합물들을 지칭하는데, 일부 화합물 쌍은 유사한 것이기에 선으로 연결되며, 다른 일부 화합물 쌍은 유사한 것이 아니기에 선으로 연결되지 않음을 알 수 있다.
- [0064] 이러한 정보, 즉 활성화 프로파일 연산 모듈(500)에서 연산된 유사도 점수와 유사 여부가 화합물 네트워크 데이터베이스(600)에 모두 저장된다.
- [0065] 탐색 모듈(700)은 구축된 화합물 네트워크 데이터베이스(600)를 활용하여, 사용자가 입력한 화합물, 즉 유사한 화합물을 검색하고자 하는 화합물(Query)과 유사한 화합물을 탐색한다.
- [0066] 즉, 어느 하나 또는 다수의 화합물이 입력 모듈(100)을 통하여 입력된 경우, 탐색 모듈(700)은 화합물 네트워크 데이터베이스(600)에서 입력된 화합물과 유사한 것으로 저장된 유사 화합물들을 모두 탐색한다.
- [0067] 예를 들어, 도 4의 (D)에서 세 개의 적색 원들이 입력된 화합물이라면, 탐색 모듈(700)은 화합물 "1", "2", "3", "4"를 유사 화합물들로 탐색한다.
- [0068] 네트워크 점수 연산 모듈(800)은, 유사 화합물들마다 유사도 점수를 통합하여 각 유사 화합물들의 네트워크 점수(network score, S)를 연산한다. 결국, 네트워크 점수는 얼마나 유사한지 여부를 판단하는 정량적 기준이 된다.
- [0069] 예를 들어, 도 4의 (D)에서 유사 화합물로 탐색된 화합물 "1", "2", "3", "4"와 관련하여, 검색하고자 하는 화합물 세 개 중 유사 여부는 각각 1개, 1개, 2개, 2개이지만, 각각의 유사도 점수는 다를 수 있으므로, 다른 네트워크 점수가 연산될 수 있다. 예를 들어, 화합물 "1", "2", "3", "4"의 네트워크 점수(S)는 각각 8점, 6점, 4

점, 10점으로 연산될 수 있다.

- [0070] 출력 모듈(900)은 네트워크 점수가 높은 순서에 따라 정렬하여, 정렬된 순서에 따라 유사 화합물이 상기 입력 모듈에 입력된 다수의 화합물과 더 유사한 화합물인 것으로 출력한다.
- [0071] 예를 들어, 도 4의 (E)에서 화합물 "1", "2", "3", "4"의 네트워크 점수(S)는 각각 8점, 6점, 4점, 10점인바, 유사한 정도는 화합물 "4"가 가장 높은 것으로 출력되며, 차례로 화합물 "1", "2", "3"이 출력된다.
- [0072] 3. 방법의 설명 - 화합물 네트워크 데이터베이스 구축
- [0073] 도 2를 참조하여, 본 발명에 따른 방법 중 화합물 네트워크 데이터베이스 구축 방법을 설명한다.
- [0074] 사용자가 입력 모듈(100)에 한 쌍의 화합물을 입력하면, 식별 모듈(200)은 입력된 한 쌍의 화합물의 식별자(ID1, ID2)를 각각 식별한다(S100).
- [0075] 다음, 바이오어세이 선택 모듈(300)이, 바이오어세이 데이터베이스에서 상기 식별자(ID1, ID2)에 해당하는 화합물이 포함된 바이오어세이들을 각각 선택한다(S110).
- [0076] 활성화 여부 확인 모듈(400)은, 기 설정된 방법을 이용하여, 상기 선택된 바이오어세이들에서 이에 대응하는 화합물에 대한 활성화 여부를 확인한다(S120).
- [0077] 활성화 프로파일 연산 모듈이, S120 단계에서 확인된 활성화 여부를 이용하여 상기 입력된 한 쌍의 화합물의 유사도 점수를 연산하고(S130), 상기 연산된 유사도 점수가 기 설정된 기준값 이상인 경우, 상기 한 쌍의 화합물의 유사한 것으로 판단함으로써, 유사 여부를 판단한다(S140).
- [0078] 이와 같이 판단된 유사 여부 및 연산된 유사도 점수가 화합물 네트워크 데이터베이스(600)에 저장된다(S150).
- [0079] 추가로 입력할 다른 한 쌍의 화합물이 있는 경우(S160), 이와 같은 과정이 다시 반복된다.
- [0080] 이와 같은 방법으로 화합물 네트워크 데이터베이스(600)가 구축되면, 여기에는 어느 하나의 화합물마다 각각 유사한 것으로 판단된 다른 하나 이상의 화합물과 비유사한 것으로 판단된 다른 하나 이상의 화합물이 저장될 수 있다.
- [0081] 4. 방법의 설명 - 약물 가상 탐색 방법
- [0082] 도 3을 참조하여, 본 발명에 따른 방법 중 약물 가상 탐색 방법을 설명한다. 여기에서, 약물 가상 탐색 방법이라 지칭함은, 본 발명에 의하여 화합물을 입력하면 유사 화합물들이 유사 정도에 따라 출력되어, 결과적으로 약물 가상 탐색에 도움을 주기 때문이다.
- [0083] 전술하였던, 화합물 네트워크 데이터베이스(600)가 구축된 이후에 수행될 수 있다.
- [0084] 사용자가 입력 모듈(100)을 통하여, 유사 화합물을 확인하고자 하는 하나 또는 다수의 화합물(Query)을 입력하면, 식별 모듈(200)은 입력된 화합물의 식별자(ID 또는 ID1~IDn)를 식별한다(S200).
- [0085] 탐색 모듈(700)은 구축되어 있는 화합물 네트워크 데이터베이스(600)에서 상기 식별자(ID)에 해당하는 화합물과 유사한 것으로 저장된 유사 화합물들을 모두 탐색한다(S210).
- [0086] 다음, 네트워크 점수 연산 모듈(800)은, 상기 유사 화합물들을 기준으로 하여 상기 식별자들(ID1~IDn)에 해당하는 화합물들 각각과의 유사도 점수를 통합함으로써, 상기 유사 화합물들에 대한 네트워크 점수(network score)를 각각 연산한다(S220).
- [0087] 출력 모듈은, 상기 유사 화합물들을 상기 연산된 네트워크 점수가 높은 순서에 따라 정렬하여, 정렬된 순서에 따라 상기 유사 화합물이 더 유사한 화합물인 것으로 출력한다(S230).
- [0088] 5. 검증실험
- [0089] 바이오어세이 데이터베이스(20)로서, 미국보건연구원에서 제공하는 대규모 생물학적 검증(bioassay) 실험 데이

터베이스인 PubChem을 사용하였으며, 바이오어세이 활성 화합물 목록을 입력 값으로 교차 검증 방식을 사용하였다.

[0090] 예측 성능은 데이터마이닝 성능 평가에 일반적으로 사용되는 ROC(Receiver Operating Characteristic) 곡선의 아래쪽 넓이를 합산한 값인 AUC를 이용하여 평가하였다.

[0091] 도 5는 사용된 데이터들을 표로서 정리한 것이며, 도 6은 본 발명에 적용하여 검정된 전체 화합물 개수의 AUC 값을 도시한 것이다. 도 6에 도시된 바와 같이, AUC 값이 0.7을 상회하며 특정 식별자에서는 1에 매우 근접함을 알 수 있다. 즉, 본 발명에 따른 방법을 이용하여 우수한 예측 성능으로서 가상 유사 화합물이 확인됨을 검증하였다.

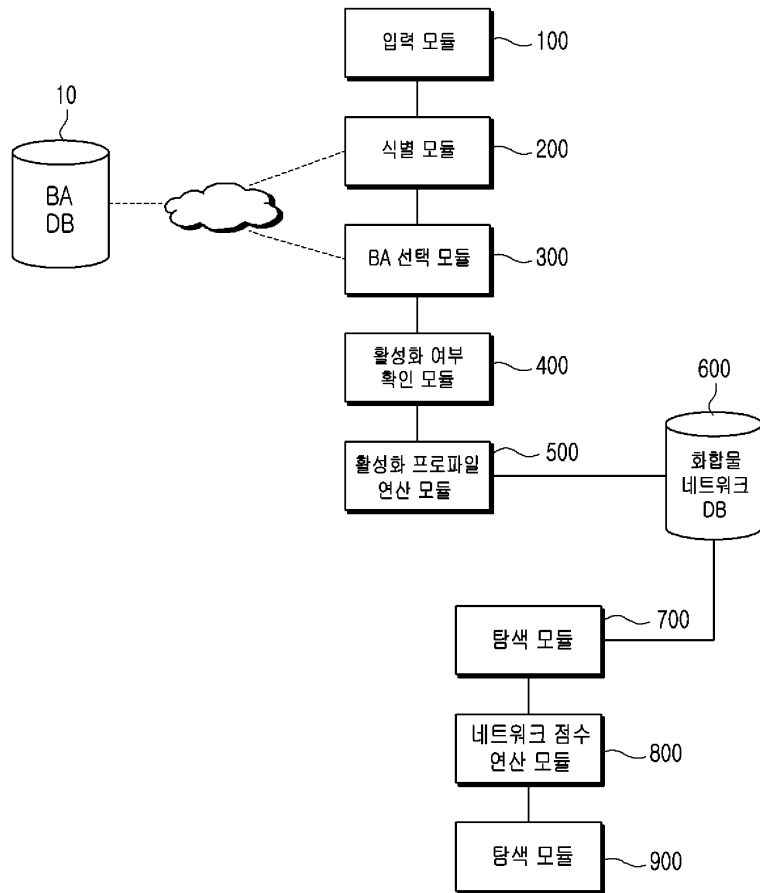
[0092] 상기에서는 본 발명의 바람직한 실시 예를 참조하여 설명하였지만, 당업계에서 통상의 지식을 가진 자라면 이하의 특허 청구범위에 기재된 본 발명의 사상 및 영역을 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

부호의 설명

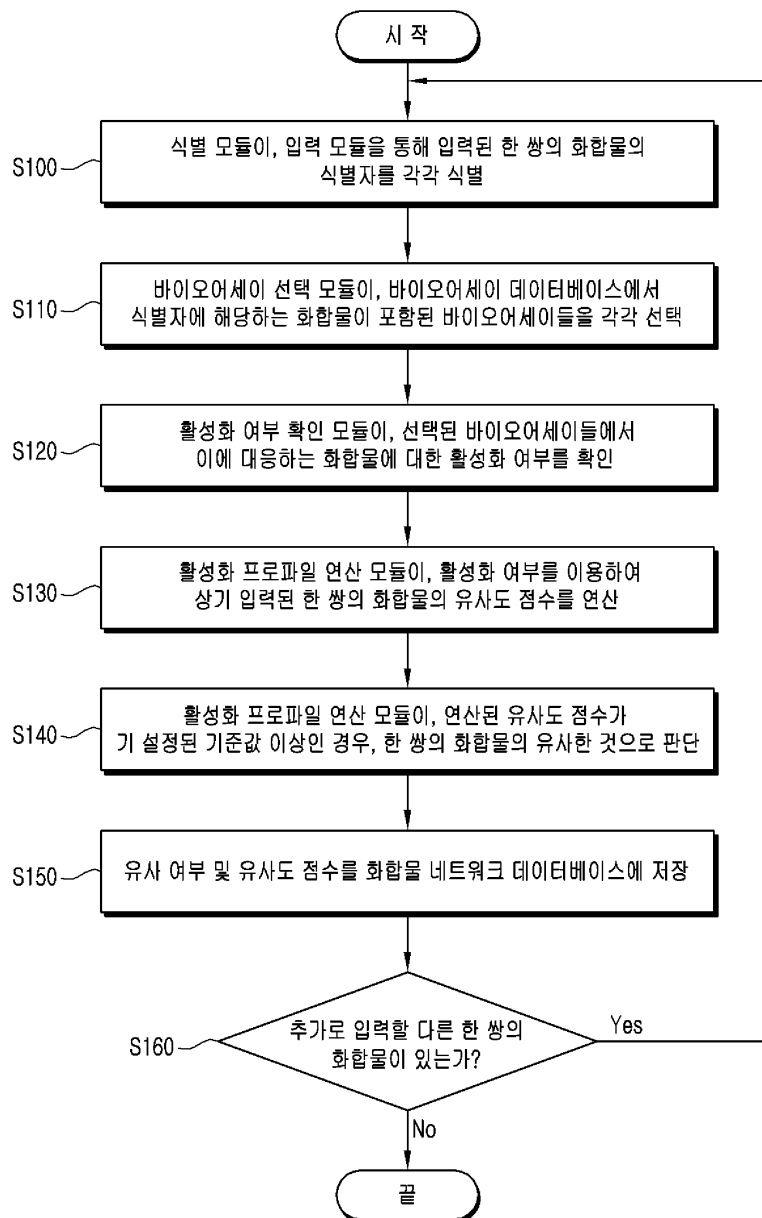
[0093] 10: 바이오어세이 데이터베이스
100: 입력 모듈
200: 식별 모듈
300: 바이오어세이 선택 모듈
400: 활성화 여부 확인 모듈
500: 활성화 프로파일 연산 모듈
600: 화합물 네트워크 데이터베이스
700: 탐색 모듈
800: 네트워크 점수 연산 모듈
900: 출력 모듈

도면

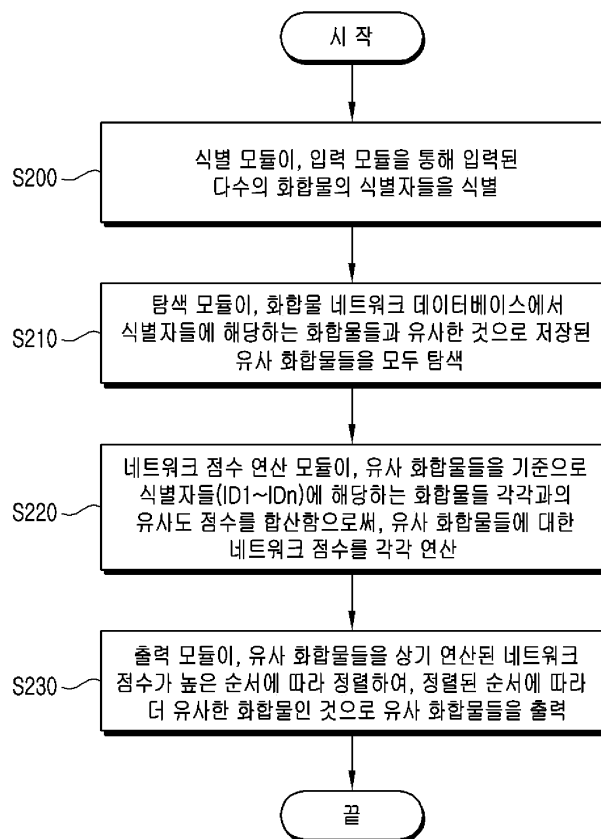
도면1



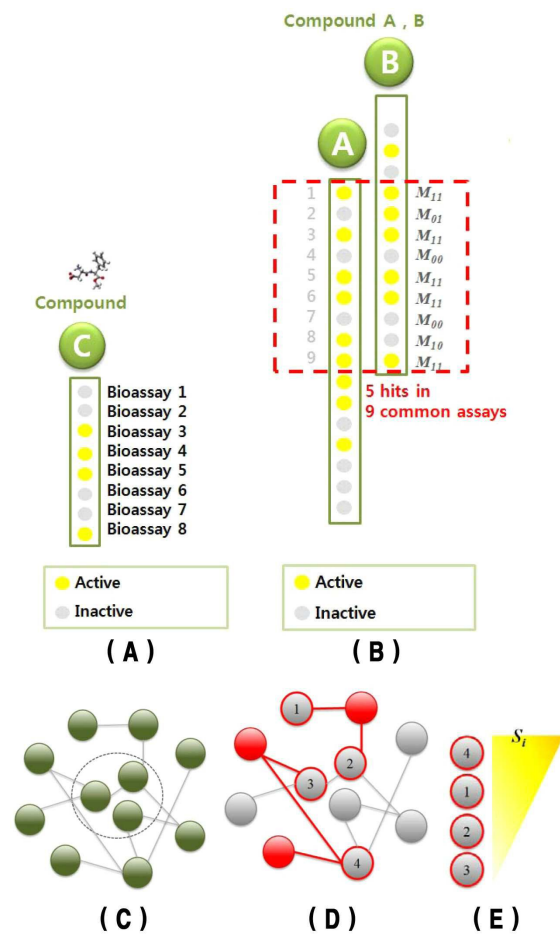
도면2



도면3



도면4



도면5

식별자	식별자 설명	조건 (타겟 단백질)	활성 화합물 개수	검정된 전체 화 합물 개수
743255	Inhibitors of USP1/UAF1	USP1/UAF1	904	389560
743269	TRFRET-based biochemical primary high throughput screening assay to identify inhibitors of HIV-1 LEDGF/p75 DNA Integration	HIV-1 LEDGF/p75 DNA Integration	2353	369939
743279	qHTS for Inhibitors of Inflammasome Signaling: IL-1-beta AlphaLISA Primary Screen	IL-1-beta AlphaLISA	17186	362038
1053174	ALPHAscreen artifact counterscreen Measured in Biochemical System Using Plate Reader - 7053-03_Inhibitor_Dose_CherryPick_Activity		141	1503
743445	Development of Small Molecule Probes of the Histone Methyltransferase, NSD2 Measured in Biochemical System Using Plate Reader - 7053-01_Inhibitor_SinglePoint_HTS_Activity_Set2	NSD2,H3K36, histone methyltransferase	1662	309676
1053136	Confirmation assay for inhibitors of LEDGF/p75-dependent integration: TR-FRET-based biochemical high throughput assay to identify inhibitors of HIV-1 LEDGF/p75 DNA Integration	HIV-1 LEDGF/p75 DNA Integration	731	2020
1053173	Development of Small Molecule Probes of the Histone Methyltransferase, NSD2 Measured in Biochemical System Using Plate Reader - 7053-01_Inhibitor_Dose_CherryPick_Activity	NSD2,H3K36, histone methyltransferase	390	1503
1053131	Counterscreen for inhibitors of LEDGF/p75-dependent integration: TR-FRET-based biochemical high throughput counterscreen assay to identify activators of HIV-1 Integrase multimerization	HIV-1 LEDGF/p75 DNA Integration	343	2020
1053124	A549 Cytotoxicity Assay Measured in Cell-Based System Using Plate Reader - 7071-06_Inhibitor_Dose_CherryPick_Activity_Set2		35	307
743266	qHTS of PTHR Inhibitors: Primary Screen	PTHR	308	405676
1117264	Fluorescence-based biochemical high throughput primary assay to identify inhibitors of Trypanosoma brucei RNA editing ligase 1 (TbREL1)	Trypanosoma brucei RNA editing ligase 1	438	89156
1053118	HepG2 Cytotoxicity Assay Measured in Cell-Based System Using Plate Reader - 7071-02_Inhibitor_Dose_CherryPick_Activity_Set5		33	231
1053119	A549 Cytotoxicity Assay Measured in Cell-Based System Using Plate Reader - 7071-06_Inhibitor_Dose_CherryPick_Activity		21	199
1053125	HepG2 Cytotoxicity Assay Measured in Cell-Based System Using Plate Reader - 7071-02_Inhibitor_Dose_CherryPick_Activity_Set6		21	307

도면6

