(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년04월24일
(11) 등록번호 10-2525028
(24) 등록일자 2023년04월19일

(51) 국제특허분류(Int. Cl.)
G16C 20/30 (2019.01) *G16C 20/70* (2019.01)
(52) CPC특허분류
G16C 20/30 (2019.02)
G16C 20/70 (2019.02)
(21) 출원번호 10-2022-0129137
(22) 출원일자 2022년10월07일
심사청구일자 2022년10월07일
(56) 선행기술조사문헌
C. Shen et al., Accuracy or novelty: what can we gain from target-specific machine-learning-based scoring functions in virtual screening?, Brief. Bioinform. Vol.22, bbaa410(2021)*
K. Lee et al., Utilizing random Forest QSAR models with optimized parameters for target identification and its application to target-fishing server, BMC Bioinformatics Vol.18, 567(2017)*
P. Durai et al., Evolutionary chemical binding similarity approach integrated with 3D-QSAR method for effective virtual screening, BMC Bioinformatics Vol.21, 309(2020)*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
충남대학교산학협력단
대전광역시 유성구 대학로 99 (궁동, 충남대학교)
(72) 발명자
강남숙
대전광역시 유성구 대학로 99 충남대학교 산학연별관 419
박경진
대전광역시 유성구 대학로159번길 13, 307호 (궁동, 백운빌라)
(74) 대리인
특허법인공룡

전체 청구항 수 : 총 6 항 심사관 : 권계민

(54) 발명의 명칭 **화합물 구조의 활성화 예측을 위한 ADis QSAR 모델 구축 장치 및 그 방법**

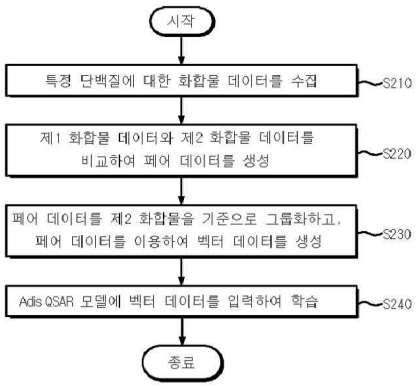
(57) 요 약

본 발명은 화합물 구조의 활성화 예측을 위한 ADis QSAR 모델 구축 장치 및 그 방법에 대한 것이다.

본 발명에 따른 화합물 구조의 활성화 예측을 위한 ADis QSAR 모델 구축 장치는 특정 단백질에 대한 활성 · 비활성 화합물 데이터를 수집하고, 수집된 화합물 데이터를 제1 그룹 및 제2 그룹으로 분류하는 데이터 수집부, 상

(뒷면에 계속)

대 표 도 - 도2



기 제1 그룹에서 추출한 화합물 데이터와 제2 그룹에서 추출한 화합물 데이터를 매칭하고, 매칭된 화합물 데이터로부터 추출된 식별자를 이용하여 페어 데이터를 획득하는 데이터 전처리부, 상기 획득한 페어 데이터에 부여된 점수들을 합산한 다음, 합산된 점수를 이용하여 벡터 데이터를 생성하는 벡터 데이터 생성부, 그리고, ADis QSAR 모델을 구축하고, 구축된 ADis QSAR 모델에 생성된 벡터 데이터를 입력하여 상기 ADis QSAR 모델로 하여금 화합물 구조의 활성화 여부를 예측하도록 학습시키는 모델 구축부를 포함한다.

명 세 서

청구범위

청구항 1

화합물 구조의 활성화 예측을 위한 ADis QSAR 모델 구축 장치에 있어서,

특정 단백질에 대한 활성 · 비활성 화합물 데이터를 수집하고, 수집된 화합물 데이터를 제1 그룹 및 제2 그룹으로 분류하는 데이터 수집부,

상기 제1 그룹에서 추출한 화합물 데이터와 상기 제2 그룹에서 추출한 화합물 데이터를 매칭하고, 매칭된 화합물 데이터로부터 추출된 식별자를 이용하여 페어 데이터를 획득하는 데이터 전처리부,

상기 획득한 페어 데이터에 부여된 점수들을 합산한 다음, 합산된 점수를 이용하여 벡터 데이터를 생성하는 벡터 데이터 생성부,

그리고, 의사결정나무(Decision Tree) 모델을 기반으로 하는 Random Forest와 XGBoost(eXtreme Gradient Boosting)를 이용하여 ADis QSAR 모델을 구축하고, 구축된 상기 ADis QSAR 모델에 생성된 벡터 데이터를 입력하여 상기 ADis QSAR 모델로 하여금 화합물 구조의 활성화 여부를 예측하도록 학습시키는 모델 구축부를 포함하며,

상기 데이터 전처리부는,

상기 제1 그룹으로 분류된 제1 화합물 데이터와 상기 제2 그룹으로 분류된 제2 화합물 데이터를 매칭하고,

상기 매칭된 제1 화합물 데이터와 제2 화합물 데이터에 포함된 핑거프린트(Fingerprint) 식별자 들을 비교하며,

상기 핑거프린트(Fingerprint) 식별자는,

상기 화합물 구조 내에 특정 하위 구조의 존재 여부를 나타내는 일련의 이진수(bit) 형태로 나타내며,

화합물에 특정 하위 구조를 포함하고 있으면 상기 식별자의 값을 1로 하고, 상기 화합물에 특정 하위 구조를 포함하고 있지 않으면 상기 식별자의 값을 0으로 하고,

상기 벡터 데이터 생성부는 상기 페어 데이터를 상기 제2 화합물을 기준으로 정렬하여 그룹화하고, 그룹화된 상기 페어 데이터에 부여된 비교 점수를 합산하여 새로운 벡터 데이터를 생성하는 과정을 수행하는 것으로, Robust Scaler를 이용하여 상기 벡터 데이터를 스케일링하는 ADis QSAR 모델 구축 장치.

청구항 2

제1항에 있어서,

상기 데이터 수집부는,

특정 단백질에 대해 활성을 가진 활성 화합물과 활성을 가지지 않은 비활성 화합물에 대한 데이터를 수집하고,

상기 수집된 화합물 데이터 중에서 상기 활성 화합물의 구조를 대표할 수 있는 화합물 데이터를 상기 제1 그룹으로 분류하고, 상기 제1 그룹으로 분류되지 않은 화합물 데이터를 상기 제2 그룹으로 분류하는 ADis QSAR 모델 구축 장치.

청구항 3

제2항에 있어서,

상기 데이터 전처리부는,

상기 제1 화합물 데이터에 포함된 식별자와 상기 제2 화합물 데이터에 포함된 식별자를 상호 비교하고,

비교한 결과에 따라 점수를 부여하고, 점수가 부여된 제1 화합물 데이터-제2 화합물 데이터를 페어 데이터로 설정하는 ADis QSAR 모델 구축 장치.

청구항 4

ADis QSAR 모델 구축 장치를 이용한 ADis QSAR 모델 구축 방법에 있어서,

특정 단백질에 대한 활성 ·비활성 화합물 데이터를 수집하고, 수집된 화합물 데이터를 제1 그룹 및 제2 그룹으로 분류하는 단계,

상기 제1 그룹에서 추출한 화합물 데이터와 상기 제2 그룹에서 추출한 화합물 데이터를 매칭하고, 매칭된 화합물 데이터로부터 추출된 식별자를 이용하여 페어 데이터를 획득하는 단계,

상기 획득한 페어 데이터에 부여된 점수들을 합산한 다음, 합산된 점수를 이용하여 벡터 데이터를 생성하는 단계, 그리고,

의사결정나무(Decision Tree) 모델을 기반으로 하는 Random Forest와 XGBoost(eXtreme Gradient Boosting)를 이용하여 ADis QSAR 모델을 구축하고, 구축된 상기 ADis QSAR 모델에 생성된 벡터 데이터를 입력하여 상기 ADis QSAR 모델로 하여금 화합물 구조의 활성화 여부를 예측하도록 학습시키는 단계를 포함하며,

상기 페어 데이터를 획득하는 단계는,

상기 제1 그룹으로 분류된 제1 화합물 데이터와 상기 제2 그룹으로 분류된 제2 화합물 데이터를 매칭하고,

상기 매칭된 제1 화합물 데이터와 제2 화합물 데이터에 포함된 핑거프린트(Fingerprint) 식별자 들을 비교하며,

상기 핑거프린트(Fingerprint) 식별자는,

상기 화합물 구조 내에 특정 하위 구조의 존재 여부를 나타내는 일련의 이진수(bit) 형태로 나타내며,

화합물에 특정 하위 구조를 포함하고 있으면 식별자의 값을 1로 하고, 화합물에 특정 하위 구조를 포함하고 있지 않으면 식별자의 값을 0으로 하고,

상기 벡터 데이터 생성하는 단계는 상기 페어 데이터를 상기 제2 화합물을 기준으로 정렬하여 그룹화하고, 그룹화된 상기 페어 데이터에 부여된 비교 점수를 합산하여 새로운 벡터 데이터를 생성하는 과정을 수행하는 것으로, Robust Scaler를 이용하여 상기 벡터 데이터를 스케일링하는 ADis QSAR 모델 구축 방법.

청구항 5

제4항에 있어서,

상기 화합물 데이터를 제1 그룹 및 제2 그룹으로 분류하는 단계는,

특정 단백질에 대해 활성을 가진 활성 화합물과 활성을 가지지 않은 비활성 화합물에 대한 데이터를 수집하고,

상기 수집된 화합물 데이터 중에서 활성 화합물의 구조를 대표할 수 있는 화합물 데이터를 상기 제1 그룹으로 분류하고, 상기 제1 그룹으로 분류되지 않은 화합물 데이터를 상기 제2 그룹으로 분류하는 ADis QSAR 모델 구축 방법.

청구항 6

제5항에 있어서,

상기 페어 데이터를 획득하는 단계는,

상기 제1 화합물 데이터에 포함된 식별자와 상기 제2 화합물 데이터에 포함된 식별자를 상호 비교하고,

비교한 결과에 따라 점수를 부여하고, 점수가 부여된 제1 화합물 데이터-제2 화합물 데이터를 페어 데이터로 설정하는 ADis QSAR 모델 구축 방법

청구항 7

삭제

청구항 8

삭제

청구항 9

삭제

청구항 10

삭제

청구항 11

삭제

청구항 12

삭제

청구항 13

삭제

청구항 14

삭제

발명의 설명

기 술 분 야

[0001] 본 발명은 화합물 구조의 활성화 예측을 위한 ADis QSAR 모델 구축 장치 및 그 방법에 관한 것으로서, 더욱 상세하게는 ADis QSAR 모델을 이용하여 활성-비활성 화합물 쌍 간의 구조와 활성 차이를 예측한 결과를 통해 신규 화합물들의 활성 상관관계를 파악하는 ADis QSAR 모델 구축 장치 및 그 방법에 관한 것이다.

배 경 기 술

[0002] 구조-활성의 정량적 관계 (Quantitative structure-activity relationship, QSAR)모델은 화학, 생물학, 공학에서 사용되는 회귀 또는 분류모델로, 화학구조와 예측하고자 하는 활성 간의 정량적인 수학적 모델을 나타낸다.

[0003] QSAR 회귀모델은 수치형태의 반응변수(Y), QSAR 분류모델은 범주형태의 반응변수와 예측변수(X)의 관계를 나타낸 것이다.

[0004] 또한, QSAR 모델링에서 예측변수는 물리화학적 특성값이나, 이론적으로 계산하는 분자 표현자를 사용하며, 반응변수는 화합물의 생물학적 활성값이 사용될 수 있다. 먼저 QSAR모델을 통해서 화학구조와 생물학적 활성과의 관계를 나타내고, 그 다음에는 QSAR모델을 이용하여 새로운 화합물의 활성을 예측한다.

[0005] 그러나, 종래의 QSAR 모델은 기본적으로 구조가 비슷한 화합물은 활성이 비슷하다는 가정에서 시작한다. 따라서, 화합물의 활성과 활성에 중요한 영향을 미치는 특징간의 상관관계를 명확하게 파악할 수 없는 문제점이 있었다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 대한민국 공개특허공보 제10-2021-0026541호(2021.03.10. 공개)

발명의 내용

해결하려는 과제

[0007] 본 발명이 이루고자 하는 기술적 과제는 ADis QSAR 모델을 이용하여 활성-비활성 화합물 쌍 간의 구조와 활성 차이를 예측하기 위한 ADis QSAR(Activity Differences Quantitative structure-activity relationship) 모델 구축 장치 및 그 방법을 제공하기 위한 것이다.

과제의 해결 수단

[0008] 이러한 기술적 과제를 이루기 위한 본 발명의 실시예에 따른 화합물 구조의 활성화 예측을 위한 ADis QSAR 모델 구축 장치는 특정 단백질에 대한 활성 · 비활성 화합물 데이터를 수집하고, 수집된 화합물 데이터를 제1 그룹 및 제2 그룹으로 분류하는 데이터 수집부, 상기 제1 그룹에서 추출한 화합물 데이터와 제2 그룹에서 추출한 화합물 데이터를 매칭하고, 매칭된 화합물 데이터로부터 추출된 식별자를 이용하여 페어 데이터를 획득하는 데이터 전처리부, 상기 획득한 페어 데이터에 부여된 점수들을 합산한 다음, 합산된 점수를 이용하여 벡터 데이터를 생성하는 벡터 데이터 생성부, 그리고, ADis QSAR 모델을 구축하고, 구축된 ADis QSAR 모델에 생성된 벡터 데이터를 입력하여 상기 ADis QSAR 모델로 하여금 화합물 구조의 활성화 여부를 예측하도록 학습시키는 모델 구축부를 포함한다.

[0009] 상기 데이터 수집부는, 특정 단백질에 대해 활성을 가진 활성 화합물과 활성을 가지지 않은 비활성 화합물에 대한 데이터를 수집하고, 상기 수집된 화합물 데이터 중에서 활성 화합물의 구조를 대표할 수 있는 화합물 데이터를 제1 그룹으로 분류하고, 제1 그룹으로 분류되지 않은 화합물 데이터를 제2 그룹으로 분류할 수 있다.

[0010] 상기 데이터 전처리부는, 상기 제1 그룹으로 분류된 제1 화합물 데이터와 상기 제2 그룹으로 분류된 제2 화합물 데이터를 매칭하고, 상기 매칭된 제1 화합물 데이터와 제2 화합물 데이터에 포함된 핑거프린트(Fingerprint) 식별자 들을 비교할 수 있다.

[0011] 상기 핑거프린트(Fingerprint) 식별자는, 화합물 구조 내에 특정 하위 구조의 존재 여부를 나타내는 일련의 이진수(bit) 형태로 나타내며, 화합물에 특정 하위 구조를 포함하고 있으면 식별자의 값을 1로 하고, 화합물에 특정 하위 구조를 포함하고 있지 않으면 식별자의 값을 0으로 할 수 있다.

[0012] 상기 데이터 전처리부는, 상기 제1 화합물 데이터에 포함된 식별자와 상기 제2 화합물 데이터에 포함된 식별자를 상호 비교하고, 비교한 결과에 따라 점수를 부여하고, 점수가 부여된 제1 화합물 데이터-제2 화합물 데이터를 페어 데이터로 설정할 수 있다.

[0013] 상기 벡터 데이터 생성부는, 상기 페어 데이터를 제2 화합물을 기준으로 정렬하여 그룹화하고, 그룹화된 페어 데이터를 이용하여 벡터 데이터를 생성할 수 있다.

[0014] 상기 모델 구축부는, 의사결정나무(Decision Tree) 모델을 기반으로 하는 Random Forest와 XGBoost(eXtreme Gradient Boosting)를 이용하여 ADis QSAR 모델을 구축할 수 있다.

[0015] 또한, 본 발명의 실시예에 따른 ADis QSAR 모델 구축 장치를 이용한 ADis QSAR 모델 구축 방법은 특정 단백질에 대한 활성 · 비활성 화합물 데이터를 수집하고, 수집된 화합물 데이터를 제1 그룹 및 제2 그룹으로 분류하는 단계, 상기 제1 그룹에서 추출한 화합물 데이터와 제2 그룹에서 추출한 화합물 데이터를 매칭하고, 매칭된 화합물 데이터로부터 추출된 식별자를 이용하여 페어 데이터를 획득하는 단계, 상기 획득한 페어 데이터에 부여된 점수들을 합산한 다음, 합산된 점수를 이용하여 벡터 데이터를 생성하는 단계, 그리고, ADis QSAR 모델을 구축하고, 구축된 ADis QSAR 모델에 생성된 벡터 데이터를 입력하여 상기 ADis QSAR 모델로 하여금 화합물 구조의 활성화 여부를 예측하도록 학습시키는 단계를 포함한다.

발명의 효과

[0016] 이와 같이 본 발명에 따르면, 화합물의 활성과 활성에 중요한 영향을 미치는 특징간의 상관관계를 추출하여 ADis QSAR 모델을 구축하고, 구축된 ADis QSAR 모델을 이용해서 아직 합성되지 않은 화합물의 활성을 미리 예측할 수 있으며, 화합물간의 활성차이를 이진수 형태의 핑거프린트(Fingerprint)를 통해 나타내므로 다른 방법들에 대비하여 적은 활성 화합물에 대한 정보를 이용하더라도 ADis QSAR 모델의 성능을 향상시켜 화합물의 활성과 활성에 중요한 영향을 미치는 특징간의 상관관계를 명확하게 파악할 수 있다.

도면의 간단한 설명

[0017] 도 1은 본 발명의 실시예에 따른 ADis QSAR 모델 구축 장치에 대한 구성도이다.

도 2는 본 발명의 실시예에 따른 ADis QSAR 모델 구축 장치를 이용한 ADis QSAR 모델 구축 방법을 설명하기 위한 순서도이다.

도 3은 도 2에 도시된 S240단계에서 구축된 ADis QSAR 모델의 성능을 평가한 결과를 나타내는 예시도이다.

발명을 실시하기 위한 구체적인 내용

[0018] 이하 첨부된 도면을 참조하여 본 발명에 따른 바람직한 실시예를 상세히 설명하기로 한다. 이 과정에서 도면에 도시된 선들의 두께나 구성요소의 크기 등은 설명의 명료성과 편의상 과장되게 도시되어 있을 수 있다.

[0019] 또한 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서, 이는 측정 대상자, 운용자의 의도 또는 관례에 따라 달라질 수 있다. 그러므로 이러한 용어들에 대한 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.

[0020] 이하에서는 도 1을 이용하여 본 발명의 실시예에 따른 ADis QSAR 모델 구축 장치에 대해 더욱 상세하게 설명한다.

[0021] 도 1은 본 발명의 실시예에 따른 ADis QSAR 모델 구축 장치에 대한 구성도이다.

[0022] 도 1에 도시된 바와 같이, 본 발명의 실시예에 따른 ADis QSAR 모델 구축 장치(100)는 데이터 수집부(110), 데이터 전처리부(120), 백터 데이터 생성부(130) 및 모델 구축부(140)를 포함한다.

[0023] 먼저, 데이터 수집부(110)는 특정 단백질에 대한 활성 · 비활성 화합물 데이터를 수집하고, 수집된 화합물 데이터를 복수의 그룹으로 분류한다.

[0024] 데이터 전처리부(120)는 그룹화된 화합물 데이터를 상호 비교하여 화합물 구조 정보를 포함하고 있는 식별자를 추출하며, 추출된 식별자에 점수를 부여한 다음, 부여된 점수를 이용하여 페어 데이터를 획득한다.

[0025] 백터 데이터 생성부(130)는 획득한 페어 데이터를 특정 화학물을 기준으로 정렬하여 그룹을 생성하고, 생성된 그룹에 포함된 식별자의 점수들을 위치별로 합산하여 벡터 데이터를 생성한다.

[0026] 모델 구축부(140)는 ADis QSAR 모델을 구축하고, 구축된 ADis QSAR 모델에 생성된 벡터 데이터를 입력하여 훈련시킨다.

[0027] 이하에서는 도 2 및 도 3을 이용하여 본 발명의 실시예에 따른 ADis QSAR 모델 구축 장치를 이용한 ADis QSAR 모델 구축 방법에 대해 더욱 상세하게 설명한다.

[0028] 도 2는 본 발명의 실시예에 따른 ADis QSAR 모델 구축 장치를 이용한 ADis QSAR 모델 구축 방법을 설명하기 위한 순서도이다.

[0029] 도 2에 도시된 바와 같이, 본 발명의 실시예에 따른 ADis QSAR 모델 구축 장치(100)는 특정 단백질에 대한 화합물 데이터를 수집한다(S210).

[0030] 데이터 수집부(110)는 특정 단백질에 대해 활성을 가진 활성 화합물과 활성을 가지지 않은 비활성 화합물에 대한 데이터를 수집한다.

[0031] 그리고, 데이터 수집부(110)는 수집된 화합물 데이터를 복수의 그룹으로 분류한다. 부연하자면, 데이터 수집부(110)는 활성 화합물의 구조를 대표할 수 있는 화합물 데이터를 제1 그룹으로 분류하고, 제1 그룹으로 분류되지 않은 화합물 데이터를 제2 그룹으로 분류한다.

[0032] 이하에서는 발명의 설명을 편리하게 하기 위하여 제1 그룹으로 분류된 화합물 데이터를 제1 화합물 데이터로 설정하고, 제2 그룹으로 분류된 화합물 데이터를 제2 화합물 데이터로 설정한다.

[0033] S210단계가 완료되면, 데이터 전처리부(120)는 제1 화합물 데이터와 제2 화합물 데이터를 비교하여 페어 데이터를 생성한다(S220).

[0034] 데이터 전처리부(120)는 제1 화합물 데이터와 제2 화합물 데이터를 매칭한다. 그리고, 데이터 전처리부(120)는 매칭된 제1 화합물 데이터와 제2 화합물 데이터에 포함된 핑거프린트(Fingerprint) 식별자 들을 비교한다.

[0035] 이를 다시 설명하면, 화합물에는 구조적 특징들을 표현해 줄 수 있는 값, 즉 표현자를 포함하며, 표현자는 원자

개수, 분자량, 결합의 종류 등을 포함한다.

[0036] 한편, 표현자는 화합물 구조 내에 특정 하위 구조의 존재 여부를 나타내는 일련의 이진수(bit) 형태의 핑거프린트(Fingerprint)를 이용하여 표현된다.

[0037] 핑거프린트(Fingerprint)는 Morgan fingerprint 중에서 Extended connectivity fingerprint(ECFP)를 사용하며, 이는 화합물의 각 원자를 탐색하면서 원자를 기준으로 특정 반경(Radius)으로 연결되어 있는 구조들의 모든 경로를 암호화한다. 각 고유 경로는 설정된 비트 수를 기준으로 미리 정의된 식별자(Integer Identifier)에 데이터 매핑된다.

[0038] 자세히는 화합물에 특정 하위 구조를 포함하고 있으면 식별자의 값을 1로 하고, 화합물에 특정 하위 구조를 포함하고 있지 않으면 식별자의 값을 0으로 한다.

[0039] 그 다음, 데이터 전처리부(120)는 제1 화합물 데이터에 포함된 식별자와 제2 화합물 데이터에 포함된 식별자를 상호 비교한다. 그리고, 데이터 전처리부(120)는 비교한 결과에 따라 점수를 부여한다.

[0040] 제1 화합물 데이터의 식별자와 제2 화합물 데이터의 식별자 값이 모두 0이면 데이터 전처리부(120)는 비교 점수를 0점으로 부여하고, 제1 화합물 데이터의 식별자 값은 0이고 제2 화합물 데이터의 식별자 값은 1이면, 데이터 전처리부(120)는 비교 점수를 1점으로 부여한다.

[0041] 또한, 제1 화합물 데이터의 식별자 값은 1이고 제2 화합물 데이터의 식별자 값은 0이면, 데이터 전처리부(120)는 비교 점수를 2점으로 부여한다. 마지막으로 제1 화합물 데이터의 식별자와 제2 화합물 데이터의 식별자 값이 모두 1이면 데이터 전처리부(120)는 비교 점수를 3점으로 부여한다.

[0042] 그 다음, 데이터 전처리부(120)는 점수가 부여된 제1 화합물 데이터-제2 화합물 데이터를 페어 데이터로 설정한다.

[0043] S220단계가 완료되면, 백터 데이터 생성부(130)는 페어 데이터를 제2 화합물을 기준으로 정렬하여 그룹화하고, 그룹화된 페어 데이터를 이용하여 벡터 데이터를 생성한다(S230).

[0044] 부연하자면, 벡터 데이터 생성부(130)는 페어 데이터를 제2 화합물을 기준으로 정렬하여 그룹화한다. 그러면, 제2 화합물 하나에 대응하여 제1 화합물의 개수만큼 페어 데이터가 존재하게 된다.

[0045] 그 다음, 벡터 데이터 생성부(130)는 그룹화된 페어 데이터에 부여된 비교 점수를 합산하여 새로운 벡터 데이터를 생성하는 과정을 수행한다. 이때, 벡터 데이터 생성부(130)는 Robust Scaler를 이용하여 벡터 데이터를 스케일링한다.

[0046] 모델 구축부(140)는 ADis QSAR 모델을 구축하고, 구축된 ADis QSAR 모델에 벡터 데이터를 입력하여 학습시킨다(S240).

[0047] 모델 구축부(140)는 의사결정나무(Decision Tree) 모델을 기반으로 만들어진 Random Forest와 XGBoost(eXtreme Gradient Boosting)를 이용하여 ADis QSAR 모델을 구축한다.

[0048] 그 다음, 모델 구축부(140)는 구축된 ADis QSAR 모델에 벡터 데이터를 입력하여 학습시킨다. ADis QSAR 모델에 대한 학습이 완료되면, 모델 구축부(140)는 ADis QSAR 모델에 대한 성능을 평가한다.

[0049] 본 발명에서는 ADis QSAR 모델의 성능 측정을 위해 IRAK4 (Interleukin-1 receptor-associated kinase 4) target에 대해 pIC_{50} 값이 8 이상의 활성을 가진 활성 화합물과 pIC_{50} 값이 8 미만의 비활성 화합물을 데이터로 사용하였다.

[0050] 이를 다시 설명하면, 모델 구축부(140)는 ChEMBL database로부터 획득한 화합물 데이터를 트레이닝 데이터 셋(활성 327 + 비활성 409)과 테스트 데이터 셋(활성 322 + 비활성 393)으로 분류한다. 그리고, 모델 구축부(140)는 트레이닝 데이터 셋을 제1 그룹(활성 52)과 제2 그룹(활성 275 + 비활성 409)으로 분류하고, 테스트 셋은 제1 그룹 없이 제2 그룹(활성 322 + 비활성 393)으로 생성한다.

[0051] 도 3은 도 2에 도시된 S240단계에서 구축된 ADis QSAR 모델의 성능을 평가한 결과를 나타내는 예시도이다.

[0052] 도 3에 도시된 바와 같이, Random Forest 학습을 통한 모델의 성능은 Train (Precision : 0.79, Recall 0.86, F1-score : 0.82, Specificity : 0.84, AUC : 0.85), Test (Precision : 0.88, Recall 0.92, F1-score : 0.9, Specificity : 0.9, AUC : 0.91)의 결과를 보였으며, XGBoost 학습을 통한 모델의 성능은 Train (Precision :

0.96, Recall : 0.96, F1-score 0.96, Specificity : 0.97, AUC : 0.96), Test (Precision : 0.89, Recall : 0.91, F1-score 0.9, Specificity : 0.91, AUC : 0.91)의 결과를 보였다.

[0053] 이와 같이 본 발명에 따른 ADis QSAR 모델 구축 장치는 화합물의 활성과 활성에 중요한 영향을 미치는 특징간의 상관관계를 추출하여 ADis QSAR 모델을 구축하고, 구축된 ADis QSAR 모델을 이용해서 아직 합성되지 않은 화합물의 활성을 미리 예측할 수 있으며, 화합물간의 활성차이를 이진수 형태의 핑거프린트(Fingerprint)를 통해 나타내므로 다른 방법들에 대비하여 적은 활성 화합물에 대한 정보를 이용하더라도 ADis QSAR 모델의 성능을 향상시킬 수 있다.

[0054] 한편, 본 발명의 실시예에서는 모델의 성능 신뢰도를 증명하기 위해 추가적으로 Benchmark 데이터 셋을 제공해주는 DUDE(A Database of Useful Decoys : Enhanced)에서 Kinase target 11개에 대한 데이터를 수집하여 target 데이터 별로 모델을 훈련시키고 예측 성능을 평가할 수도 있다.

[0055] 그 결과 데이터 셋별 Random Forest 학습을 통한 모델의 성능은 Train (AUC : 0.82 ~ 0.95), Test (AUC : 0.61 ~ 0.91) 사이의 결과를 보였으며, XGBoost 학습을 통한 모델의 성능은 Train (AUC : 0.97 ~ 1), Test (AUC : 0.84 ~ 0.96) 사이의 결과를 보였다.

[0056] 본 발명은 도면에 도시된 실시예를 참고로 하여 설명되었으나 이는 예시적인 것에 불과하며, 당해 기술이 속하는 분야에서 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 따라서 본 발명의 진정한 기술적 보호범위는 아래의 특허청구범위의 기술적 사상에 의하여 정해져야 할 것이다.

부호의 설명

[0057] 100 : ADis QSAR 모델 구축 장치

110 : 데이터 수집부

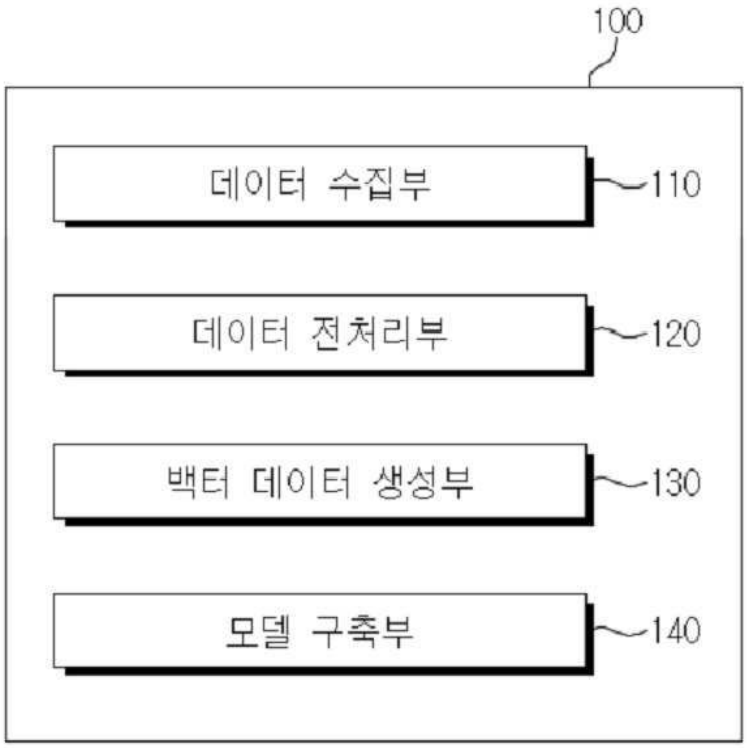
120 : 데이터 전처리부

130 : 백터 데이터 생성부

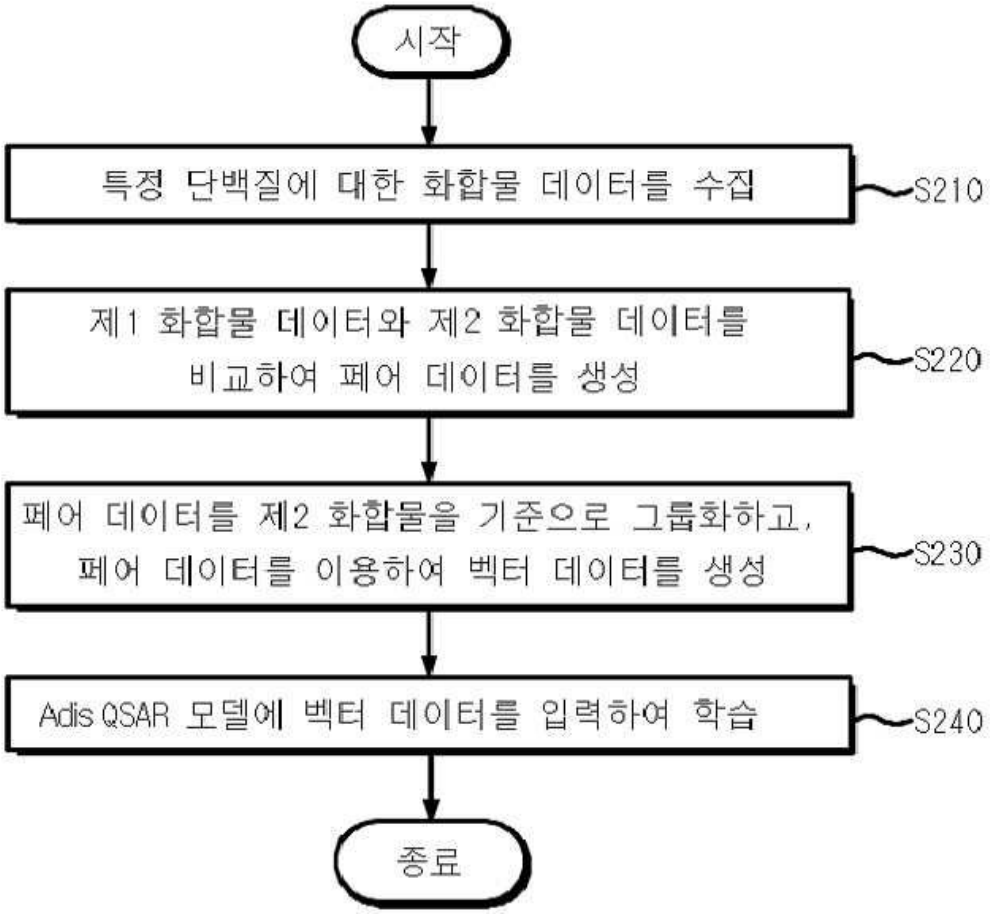
140 : 모델 구축부

도면

도면1



도면2


시작
특정 단백질에 대한 화합물 데이터를 수집
S210
제1 화합물 데이터와 제2 화합물 데이터를 비교하여 페어 데이터를 생성
S220
페어 데이터를 제2 화합물을 기준으로 그룹화하고, 페어 데이터를 이용하여 벡터 데이터를 생성
S230
Adis QSAR 모델에 벡터 데이터를 입력하여 학습
S240
종료

도면3

< Random Forest 방법>

Target	ECFP(2) nbits(256)	ECFP(2) nbits(512)
ABL1	train(0.85), test(0.8)	train(0.86), test(0.82)
AKT1	train(0.95), test(0.88)	train(0.94), test(0.88)
AKT2	train(0.86), test(0.81)	train(0.9), test(0.83)
BRAF	train(0.94), test(0.71)	train(0.95), test(0.8)
CDK2	train(0.84), test(0.75)	train(0.82), test(0.77)
CSF1R	train(0.85), test(0.72)	train(0.83), test(0.76)
EGFR	train(0.93), test(0.89)	train(0.93), test(0.91)
FAK1	train(0.95), test(0.87)	train(0.93), test(0.91)
IGF1R	train(0.93), test(0.78)	train(0.93), test(0.8)
JAK2	train(0.9), test(0.72)	train(0.88), test(0.76)
KIT	train(0.86), test(0.81)	train(0.82), test(0.65)

Target	ECFP(4) nbits(256)	ECFP(4) nbits(512)
ABL1	train(0.88), test(0.8)	train(0.87), test(0.82)
AKT1	train(0.94), test(0.87)	train(0.95), test(0.89)
AKT2	train(0.83), test(0.81)	train(0.76), test(0.78)
BRAF	train(0.95), test(0.89)	train(0.95), test(0.82)
CDK2	train(0.84), test(0.75)	train(0.82), test(0.78)
CSF1R	train(0.84), test(0.71)	train(0.83), test(0.76)
EGFR	train(0.91), test(0.89)	train(0.93), test(0.9)
FAK1	train(0.95), test(0.89)	train(0.93), test(0.91)
IGF1R	train(0.94), test(0.78)	train(0.93), test(0.8)
JAK2	train(0.9), test(0.73)	train(0.91), test(0.76)
KIT	train(0.83), test(0.82)	train(0.84), test(0.65)

< XGBoost 방법>

Target	ECFP(2) nbits(256)	ECFP(2) nbits(512)
ABL1	train(1), test(0.83)	train(1), test(0.92)
AKT1	train(1), test(0.96)	train(1), test(0.96)
AKT2	train(1), test(0.89)	train(1), test(0.93)
BRAF	train(0.99), test(0.89)	train(0.97), test(0.91)
CDK2	train(1), test(0.92)	train(1), test(0.94)
CSF1R	train(1), test(0.87)	train(0.99), test(0.89)
EGFR	train(1), test(0.96)	train(1), test(0.97)
FAK1	train(1), test(0.96)	train(1), test(0.96)
IGF1R	train(1), test(0.9)	train(1), test(0.91)
JAK2	train(1), test(0.89)	train(1), test(0.9)
KIT	train(0.96), test(0.84)	train(1), test(0.87)

Target	ECFP(4) nbits(256)	ECFP(4) nbits(512)
ABL1	train(1), test(0.88)	train(0.99), test(0.9)
AKT1	train(1), test(0.96)	train(1), test(0.96)
AKT2	train(1), test(0.89)	train(1), test(0.93)
BRAF	train(0.99), test(0.89)	train(0.97), test(0.91)
CDK2	train(1), test(0.92)	train(1), test(0.94)
CSF1R	train(1), test(0.87)	train(0.99), test(0.89)
EGFR	train(1), test(0.96)	train(1), test(0.97)
FAK1	train(1), test(0.96)	train(1), test(0.96)
IGF1R	train(1), test(0.9)	train(1), test(0.91)
JAK2	train(1), test(0.89)	train(1), test(0.9)
KIT	train(0.96), test(0.84)	train(1), test(0.87)