(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일
2020 년 9 월 17 일 (17.09.2020)

WIPO | PCT

(10) 국제공개번호
WO 2020/184816 A1

(51) 국제특허분류:
G16H 70/40 (2018.01) *G16C 20/70* (2019.01)
G16H 50/70 (2018.01) *G16C 60/00* (2019.01)

(21) 국제출원번호: PCT/KR2019/017793

(22) 국제출원일: 2019 년 12 월 16 일 (16.12.2019)

(25) 출원언어: 한국어

(26) 공개언어: 한국어

(30) 우선권정보:
10-2019-0028788 2019 년 3 월 13 일 (13.03.2019) KR
10-2019-0028789 2019 년 3 월 13 일 (13.03.2019) KR
PCT/KR2019/002918 2019 년 3 월 13 일 (13.03.2019) KR
PCT/KR2019/002919 2019 년 3 월 13 일 (13.03.2019) KR
10-2019-0163398 2019 년 12 월 10 일 (10.12.2019) KR

(71) 출원인: 주식회사 메디리타 (**MEDIRITA**) [KR/KR]; 08389 서울시 구로구 디지털로30길 28, 8층 803호 (구로동, 마리오타워), Seoul (KR).

(72) 발명자: 배영우 (**PAE, Young Woo**); 04724 서울시 성동구 동호로 100, 101동 805호, Seoul (KR). 진승현 (**JIN, Seung-Hyun**); 03907 서울시 마포구 상암산로1길 24, 410동 1602호, Seoul (KR).

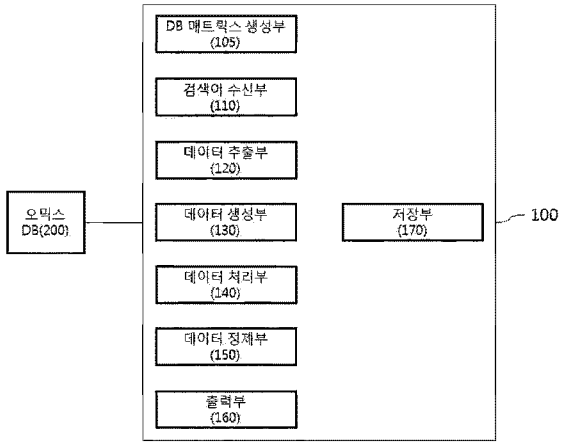
(74) 대리인: 강범주 (**KANG, Beom Ju**); 06252 서울시 강남구 역삼로 114, 8층, Seoul (KR).

(81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 역내 권리의 보호를 위하여): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 유

(54) **Title:** DATA PROCESSING METHOD FOR DERIVING NEW DRUG CANDIDATE

(54) **발명의 명칭:** 신약 후보 물질 도출을 위한 데이터 처리 방법



105 ... DB matrix generation unit
110 ... Keyword receiving unit
120 ... Data extraction unit
130 ... Data generation unit
140 ... Data processing unit
150 ... Data refinement unit
160 ... Output unit
170 ... Storage unit
200 ... Omics DB

(57) **Abstract:** A data processing method for discovering a new drug candidate, of a data processing device, comprises the steps of: generating, from an omics DB, a DB matrix comprising selected biological entities and selected correlation types; receiving a predetermined keyword; extracting, from the DB matrix, at least one biological entity related to the predetermined keyword; extracting, from the DB matrix, a correlation between the predetermined keyword and the at least one biological entity; building a first knowledge network in which a plurality of nodes, comprising the predetermined keyword and the at least one biological entity, are connected in accordance with the correlation; calculating a graph theory indicator of the first knowledge network; and building a second knowledge network by using some nodes, among the plurality of nodes, extracted using the graph theory indicator.

(57) **요약서:** 데이터 처리 장치의 신약 후보 물질 발굴을 위한 데이터 처리 방법은 선택된 생물학적 엔티티와 선택된 상호 연관도 종류로 구성되는 DB 매트릭스를 오믹스 DB로부터 생성하는 단계; 소정의 검색어를 수신하는 단계; 상기 DB 매트릭스로부터 상기 소정의 검색어와 관련된 적어도 하나의 생물학적 엔티티를 추출하는 단계; 상기 DB 매트릭스로부터 소정의 검색어와 상기 적어도 하나의 생물학적 엔티티 간 상호 연관도를 추출하는 단계; 상기 소정의 검색어와 상기 적어도 하나의 생물학적 엔티티를 포함하는 복수의 노드를 상기 상호 연관도에 따라 연결한 제1 지식 네트워크를 생성하는 단계; 상기 제1 지식 네트워크의 그래프 이론 지표를 계산하는 단계; 그리고 상기 복수의 노드 중 상기 그래프 이론 지표를 이용하여 추출된 일부 노드를 이용하여 제2 지식 네트워크를 생성하는 단계를 포함한다.

WO 2020/184816 A1

[다음 쪽 계속]

럼 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

공개:

— 국제조사보고서와 함께 (조약 제21조(3))

명세서

발명의 명칭: 신약 후보 물질 도출을 위한 데이터 처리 방법

기술분야

[1] 본 발명은 신약 개발 방법에 관한 것으로, 보다 구체적으로는 인체 오믹스 데이터베이스(OMICS Database)로부터 신약 후보 물질을 도출하기 위한 데이터 처리 방법에 관한 것이다.

배경기술

[2] 하나의 신약을 개발하기 위하여 평균적으로 총 15년의 기간이 소요되며, 2~3조원의 비용이 발생하는 것으로 알려져 있다. 이 중에서도 전임상(preclinical trial) 이전의 신약 후보 물질을 발굴하기 위하여 약 6년의 기간이 소요되는 것으로 알려져 있다.

[3] 일반적으로, 신약을 개발하기 위한 파이프라인의 첫 단계인 신약 후보 물질을 발굴하기 위하여, 다수의 전문 연구 인력들이 막대한 양의 정보를 일일이 탐색하고, 이로부터 주요한 생물학적 엔티티(entity) 간의 연관성을 추론하는 과정을 거치고 있다.

[4] 한편, 최근 일본에서 출범된 라이프 인텔리전스 컨소시엄(Life Intelligence Consortium, 2017)에 따르면, 신약 개발에 인공지능 기술을 활용할 경우, 신약을 개발하기 위하여 소요되는 기간은 약 40% 수준으로 단축될 수 있고, 비용은 약 50% 수준으로 절감될 수 있는 것으로 예측되고 있다.

발명의 상세한 설명

기술적 과제

[5] 본 발명이 해결하고자 하는 기술적 과제는 신약 후보 물질 발굴을 위한 데이터 처리 방법을 제공하는 것이다. 본 발명이 해결하고자 하는 다른 기술적 과제는 인체 오믹스 데이터베이스(DB)로부터 계층 구조를 가지는 멀티오믹스 네트워크를 생성하고, 멀티오믹스 네트워크로부터 정제된 지식 네트워크를 생성하는 방법에 관한 것이다.

발명의 효과

[6] 신약 후보 물질 발굴을 위하여 막대한 양의 정보를 일일이 탐색하지 않고도, 소정의 검색어와 관련된 생물학적 엔티티 및 이들의 상호 연관도에 관한 정제된 정보를 단시간 내에 추출할 수 있다. 이에 따라, 신약 후보 물질 또는 신약 후보 물질의 타겟을 발굴하는데 소요되는 비용 및 기간을 현저히 줄일 수 있다.

도면의 간단한 설명

[7] 도1은 일 실시 예에 따라, 신약 후보 물질 발굴을 위한 데이터 처리 장치의 블록도이다.

[8] 도2는 일 실시 예에 따라, 데이터 처리 장치의 신약 후보 물질 발굴을 위한 데이터 처리 방법의 흐름도를 나타낸다.

[9] 도3은 일 실시 예에 따라, 입력되는 소정의 검색어를 나타낸다.

[10] 도4는 일 실시 예에 따라, 단계 S205 에서 생성된 DB매트릭스를 나타낸다.

[11] 도5는 일 실시 예에 따라, 단계 S205 에서 생성된 DB매트릭스를 나타낸다.

[12] 도6은 일 실시 예에 따른 제1 지식 네트워크이다.

[13] 도7은 일 실시 예에 따라, Participation coefficient(PC)에 따라 허브의 종류를 구분하는 것을 나타낸다.

[14] 도8은 일 실시 예에 따라, 검색어 "epilepsy syndrome"로부터 생성된 제2 지식 네트워크이다.

[15] 도9는 일 실시 예에 따라, 오믹스 레벨(생물학적 엔티티)이 입력되는 예를 나타낸다.

[16] 도10은 일 실시 예에 따라, 상호 연관도 종류가 입력되는 예를 나타낸다.

[17] 도11은 추가적 실시 예에 따라, 신약 후보 물질 발굴을 위한 데이터 처리 장치의 블록도이다.

[18] 도12는 추가적 실시 예에 따라, 데이터 처리 장치의 신약 후보 물질 발굴을 위한 데이터 처리 방법의 흐름도를 나타낸다.

[19] 도13은 일 실시 예에 따라, 데이터 처리 장치가 약물 가능 경로를 탐색하는 방법의 흐름도를 나타낸다.

발명의 실시를 위한 최선의 형태

[20] 데이터 처리 장치에서 수행되는 신약 후보 물질 발굴을 위한 데이터 처리 방법에 있어서, 선택된 생물학적 엔티티와 선택된 상호 연관도 종류로 구성되는 DB 매트릭스를 오믹스 DB로부터 생성하는 단계, 검색어를 수신하는 단계, 상기 DB 매트릭스로부터 상기 검색어와 다른 오믹스 레벨에 속하고 상기 검색어와 관련된 생물학적 엔티티들을 추출하는 단계, 상기 DB 매트릭스로부터 상기 검색어와 상기 생물학적 엔티티들 간의 상호 연관도를 추출하는 단계, 상기 검색어와 상기 생물학적 엔티티들 각각을 노드로 하고, 상기 검색어와 상기 생물학적 엔티티들 사이의 상호 연관도 또는 상기 생물학적 엔티티들 간 상호 연관도에 따라 연결선을 이용하여 복수의 노드들을 연결한 제1지식 네트워크를 생성하는 단계, 상기 제1지식 네트워크의 복수의 노드들 각각에 대해 그래프 이론 지표를 계산하는 단계, 및 상기 제1지식 네트워크의 복수의 노드들 중 상기 그래프 이론 지표를 이용하여 선택된 일부 노드들을 이용하여 제2지식 네트워크를 생성하는 단계를 포함하고, 상기 검색어는 유전자명, 단백질명, 신진대사체명, 증상명, 질환명, 화합물명 및 약품명 중 적어도 하나를 포함하고, 상기 생물학적 엔티티는 유전자, 단백질, 신진대사체, 증상, 질환, 화합물 및 약품 중 적어도 하나를 포함하며, 상기 상호 연관도의 범주는 참여(participate), 공변(covariate), 조절(regulate), 연관(associate), 결합(bind), 업레귤레이트(upregulate), 유사(resemble), 치료(treat), 다운레귤레이트(downregulates), 완화(palliate), 포함(include), 및 표출(express)을

포함하며, 상기 그래프 이론 지표는 상기 제1지식 네트워크를 구성하는 복수의 노드들 중 적어도 하나에 대한 노드 간 최단 경로, 노드 별 클러스터링 계수, 노드 별 센트럴리티 계수를 포함하고, 상기 연결선이 나타내는 상호연관도의 범주에 따라 상기 연결선의 가중치가 다르게 설정되고, 상기 노드 간 최단 경로는 상기 설정된 가중치를 반영하여 산출되고, 상기 제2지식 네트워크를 생성하는 단계는, 상기 제1지식 네트워크를 구성하는 복수의 노드들 각각에 대해 상기 노드 간 최단 경로, 상기 노드 별 클러스터링 계수, 및 상기 노드 별 센트럴리티 계수 중 적어도 하나에 대한 표준 점수를 계산하고, 상기 표준 점수가 임계 값 미만인 노드와 상기 임계 값 미만인 노드의 연결선을 삭제함으로써 상기 제2지식 네트워크를 생성하고, 상기 표준 점수는 제1 지식 네트워크를 구성하는 각 노드에 대한 소정의 그래프 이론 지표의 지표값과 제1 지식 네트워크를 구성하는 복수의 노드에 대한 그래프 이론 지표의 평균 지표값 간의 차를 표준 에러로 나눈 값이고, 상기 DB 매트릭스는, 상기 선택된 생물학적 엔티티들이 가로축 및 세로축 각각에 배치되며, 가로축과 세로축이 교차하는 지점에 상기 상호 연관도 종류가 표시되도록 생성될 수 있다.

[21] 상기 제2지식 네트워크를 생성하는 단계는, 상기 제1지식 네트워크를 구성하는 전체 연결선을 임의로 섞은 다음 상기 제1지식 네트워크의 노드들 각각에 대해 상기 표준 점수를 계산하는 단계를 포함하고, 상기 임의로 섞는 회수는 1000회 이상일 수 있다.

[22] 상기 제2지식 네트워크를 생성하는 단계는, 상기 제1지식 네트워크를 구성하는 노드들 중에서 연결선이 하나인 노드를 삭제하는 단계, 및 상기 제1지식 네트워크를 구성하는 노드들 중에서 클러스터링 계수가 0인 노드를 삭제하는 단계를 더 포함할 수 있다.

[23] 상기 상호연관도의 범주는, 상호작용(interact), 원인(cause), 발현(present), 및 위치(localize) 중 적어도 하나를 더 포함할 수 있다.

[24] 상기 제2지식 네트워크로부터 약물 가능 경로를 추출하는 단계를 더 포함하고, 상기 약물 가능 경로를 추출하는 단계는, 상기 제2지식 네트워크에 존재하는 약물-질환 노드들 각각에 대한 근접도의 표준 점수가 기준 값보다 작은 약물-질환 노드 페어들을 선택하는 단계, 상기 선택된 약물-질환 노드 페어들에 대한 경로들 중에서, 상기 경로들 각각에 존재하는 중간 노드가 기준 개수 이상인 경로들을 추출하는 단계, 및 상기 추출된 경로들 중에서, 상기 추출된 경로들의 중간 노드들의 센트렐리티 계수의 총합이 기준 값 이상인 경로를 상기 약물 가능 경로로서 추출하는 단계를 포함할 수 있다.

[25] 상기 데이터 처리 방법을 컴퓨터에서 실행시키기 위한 프로그램이 기록된 기록매체가 제공될 수 있다.

발명의 실시를 위한 형태

[26] 아래에서, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자들(이하,

통상의 기술자들)이 본 발명을 용이하게 실시할 수 있도록, 첨부되는 도면들을 참조하여 몇몇 실시 예가 명확하고 상세하게 설명될 것이다.

[27] 또한, 명세서에서 사용되는 "부" 이라는 용어는 FPGA(Field Programmable Gate Array) 또는 ASIC(Application Specific Integrated Circuit)과 같은 하드웨어 구성요소 또는 회로를 의미할 수 있다.

[28] 도1은 일 실시예에 따른 신약 후보 물질 발굴을 위한 데이터 처리 장치의 블록도이고, 도2는 일 실시 예에 따른 데이터 처리 장치의 신약 후보 물질 발굴을 위한 데이터 처리 방법의 흐름도를 나타낸다.

[29] 도1을 참조하면, 신약 후보 물질 발굴을 위한 데이터 처리 장치(100)는 DB 매트릭스 생성부(105), 검색어 수신부(110), 데이터 추출부(120), 데이터 생성부(130), 데이터 처리부(140), 데이터 정제부(150), 출력부(160), 및 저장부(170)를 포함할 수 있다. 데이터 처리 장치(100)는 적어도 하나의 컴퓨팅 장치를 포함할 수 있다. 예를 들어, 데이터 처리 장치(100)는 적어도 하나의 프로세서와 적어도 하나의 메모리를 포함할 수 있다.

[30] 도1 내지 2를 참조하면, DB 매트릭스 생성부(105)는 오믹스 DB(200)로부터 적어도 일부의 오믹스 레벨(생물학적 엔티티)들에 관한 DB 및 적어도 일부의 상호 연관도 종류들에 관한 DB로 구성되는 DB 매트릭스를 생성할 수 있다(S205). DB 매트릭스를 생성하기 위한 오믹스 레벨(생물학적 엔티티)들과 상호 연관도 종류들은 사용자에 의해 선택될 수 있다. 이를 위해, DB 매트릭스 생성부(105)는 DB 매트릭스를 생성하기 위해, 오믹스를 이루는 복수의 레벨 중 적어도 일부의 오믹스 레벨(생물학적 엔티티)을 입력 받고, 오믹스를 이루는 복수의 상호 연관도 종류 중 적어도 일부의 상호 연관도 종류를 입력 받을 수 있다.

[31] 오믹스(omics)는 체학이라고도 하며, 예를 들어 유전자체학, 전사체학, 단백질체학, 신진대사체학, 후성유전체학, 지질체학 등이 있고, 세부적으로 해부학적 구조(anatomy), 생물학적 경로(biological process), 전도경로(pathway), 약리학적 계층(pharmacological class), 증상, 질환, 화합물, 약물, 부작용 등에 관련된 내용을 포함할 수 있으나, 이로 제한되는 것은 아니다. 복수의 오믹스 레벨은 유전자 레벨, 전사 레벨, 단백질 레벨, 신진대사체 레벨, 후성유전자 레벨, 지질 레벨, 해부학적 구조 레벨, 생물학적 경로 레벨, 전도경로 레벨, 약리학적 계층레벨, 증상 레벨, 질환 레벨, 화합물 레벨, 약물 레벨 및 부작용 레벨 등을 포함할 수 있으나, 이로 제한되는 것은 아니다. 여기서, 해부학적 구조는 조직(tissue), 기관(organ) 등을 의미할 수 있고, 생물학적 경로는 세포 내 구조의 레벨에서의 위치와 같은 세포 구성성분, 유전자 온톨로지로부터 추출된 분자 기능을 포함하는 일련의 이벤트일 수 있으며, 약리학적 계층은 약리학적 효과, 작용의 메커니즘일 수 있다.

[32] 복수의 상호 연관도 종류는 "상호작용(interact)", "참여(participate)", "공변(covariate)", "조절(regulate)", "연관(associate)", "결합(bind)",

"업레귤레이트(upregulate)", "원인(cause)", "유사(resemble)", "치료(treat)", "다운레귤레이트(downregulates)", "완화(palliate)", "발현(present)", "위치(localize)", "포함(include)", "표출(express)", "감소(decrease)", "증가(increase)" 등을 포함할 수 있으며, 종류 별로 식별 번호 또는 식별 기호가 임의로 부여될 수 있다. 종류 별 식별 번호 또는 식별 기호는 사용자에 의하여 설정되거나, 자동으로 설정될 수 있다.

[33] 오믹스 DB(200)는 빅데이터 DB일 수 있으며, 본 발명의 실시예에 따른 데이터 처리 장치(100) 외부의 DB일 수 있고, 누구나 접근 가능하거나 소정의 조건 하에 인증 받은 자가 접근 가능한 글로벌 공공 DB일 수 있다. 오믹스 DB(200)는 오믹스 레벨(생물학적 엔티티)에 관한 정보 및 오믹스 레벨 내 생물학적 엔티티 간 상호 연관도에 관한 정보를 미리 저장할 수 있다. 예를 들어, 오믹스 DB는 오믹스 레벨 별 DB 및 상호 연관도 종류 별 DB를 포함할 수 있다.

[34] 오믹스 레벨 별 DB는, 예를 들어 유전자 DB, 전사 DB, 단백질 DB, 신진대사체 DB, 후성유전자 DB, 지질 DB, 해부학적 구조 DB, 생물학적 경로 DB, 전도경로 DB, 증상 DB, 질환 DB, 화합물 DB, 약물 DB 및 부작용 DB를 포함할 수 있다.

[35] 상호 연관도 종류 별 DB는 상호작용(interact) DB, 참여(participate) DB, 공변(covariate) DB, 조절(regulate) DB, 연관(associate) DB, 결합(bind) DB, 업레귤레이트(upregulate) DB, 원인(cause) DB, 유사(resemble) DB, 치료(treat) DB, 다운레귤레이트(downregulates) DB, 완화(palliate) DB, 발현(present) DB, 위치(localize) DB, 포함(include) DB, 표출(express) DB, 감소(decrease) DB, 증가(increase) DB 를 포함할 수 있다. 이들 DB는 하나의 빅데이터 DB로 통합하여 관리 및 운영되거나, 분산되어 관리 및 운용될 수 있다.

[36] 도9는 일 실시예에 따라, DB 매트릭스를 생성하기 위해 오믹스 레벨(생물학적 엔티티)이 입력되는 예를 나타내고, 도10은 일 실시예에 따라 DB 매트릭스를 생성하기 위해 상호 연관도 종류가 입력되는 예를 나타낸다. 도9를 참조하면, 출력부(160)를 통하여 복수의 오믹스 레벨이 선택될 수 있는 화면이 노출될 수 있으며, 복수의 오믹스 레벨 중 사용자 인터페이스를 통하여 적어도 일부의 오믹스 레벨이 선택될 수 있다. 그리고, 도10을 참조하면, 출력부(160)를 통하여 복수의 상호 연관도 종류가 선택될 수 있는 화면이 노출될 수 있으며, 복수의 상호 연관도 종류 중 사용자 인터페이스를 통하여 적어도 일부의 상호 연관도 종류가 선택될 수 있다.

[37] 도4와 도5는 DB 매트릭스의 예를 나타낸다. 만약, 사용자가 DB 매트릭스를 생성하기 위해 오믹스 DB의 모든 오믹스 레벨(생물학적 엔티티)과 모든 상호 연관도 종류를 선택한다면 DB 매트릭스는 도4과 같이 생성될 수 있다. 도4를 참조하면, 선택된 오믹스 레벨(생물학적 엔티티)들이 가로축 및 세로축 각각에 배치되며, 가로축 및 세로축이 교차하는 지점에 선택된 상호 연관도 종류들이 표시되도록 생성될 수 있다.

[38] 예를 들어, 유전자 레벨(Gene), 단백질 레벨(Protein), 지질 레벨(Lipid),

신진대사체 레벨(Metabolite), 해부학적 구조 레벨(Anatomy), 생물학적 경로 레벨(Biological Process), 세포적 기반(Cellular Component), 분자 기능 레벨(Molecular Function), 약물 레벨(Drug), 부작용 레벨(Side Effect), 질병 레벨(Disease), 약리학적 계층 레벨(Pharmacological Class), 및 증상 레벨(Symptom)이 DB 매트릭스의 가로축 및 세로축 각각에 배치될 수 있으며, 가로축과 세로축이 교차하는 지점에 상호 연관도 종류인 상호작용(interact, Int), 참여(participate, P), 공변(covariate, Co), 조절(regulate, Reg), 연관(associate,A), 결합(bind, B), 업레귤레이트(upregulate, U), 원인(cause, Ca), 유사(resemble, R), 치료(treat, T), 다운레귤레이트(downregulates, D), 완화(palliate, Pa), 발현(present, Pr), 위치(localize, L), 포함(include, Inc), 감소(decrease, Decre), 증가(increase, Incre), 전이(translation, Tr), 및 표출(express, E) 중 적어도 하나가 표시될 수 있다.

[39] 만약, 사용자가 DB 매트릭스를 생성하기 위해 DB 종류를 유전자 레벨(Gene), 약물 레벨(Drug), 질병 레벨(Disease)로 선택하고 DB 사이의 상호 연관도의 종류를 공변(Co), 조절(Reg), 업레귤레이트(U), 결합(B), 다운레귤레이트(D), 연관(A), 유사(R), 치료(T), 완화(Pa)로 선택한다면 DB 매트릭스는 도5와 같이 생성될 수 있다.

[40] 다시 도1 내지 2를 참조하면, 검색어 수신부(110)는 소정의 검색어를 수신할 수 있다(S200). 소정의 검색어는 사용자 인터페이스를 통하여 입력될 수 있으며, 유전자명, 단백질명, 신진대사체명, 증상명, 질환명, 화합물명 및 약물명 중 적어도 하나를 포함할 수 있다. 예를 들어, 사용자는 검색어 수신부(110)를 통해 Bupropion 이라는 약물을 검색어로서 입력하거나 epilepsy syndrome 라는 질환을 검색어로 입력할 수 있다. 도3은 소정의 검색어가 입력되는 예를 나타낸다. 도3을 참조하면, 출력부(160)를 통하여 소정의 검색어를 입력하기 위한 화면이 노출될 수 있으며, 사용자 인터페이스를 통하여 소정의 검색어가 입력될 수 있다. 도3은 질환명을 범주로 선택하며, 소정의 검색어로 epilepsy syndrome를 입력하는 예를 나타낸다.

[41] 다음으로, 데이터 추출부(120)는 단계 S200에서 수신된 소정의 검색어와 관련된 적어도 하나의 생물학적 엔티티(entity)를 생성한 DB 매트릭스를 이용하여 추출하며(S210), 소정의 검색어와 추출한 생물학적 엔티티 간 상호 연관도를 생성한 DB 매트릭스를 이용하여 추출할 수 있다(S220). 여기서, 생물학적 엔티티는 유전자, 단백질, 신진대사체, 증상, 질환, 화합물 및 약물 중 적어도 하나를 포함할 수 있으며, 소정의 검색어가 속한 레벨은 생물학적 엔티티가 속한 오믹스 레벨과 동일할 수도 있고, 상이할 수도 있다. 예를 들어, 도3에서 예시한 바와 같이, 소정의 검색어가 질환명인 epilepsy syndrome인 경우, 단계 S210에서 추출되는 생물학적 엔티티는 epilepsy syndrome과 연관된 유전자, epilepsy syndrome과 연관된 단백질, epilepsy syndrome과 연관된 신진대사체, epilepsy syndrome과 연관된 증상, epilepsy syndrome과 연관된 질환, epilepsy syndrome과 연관된 화합물 및 epilepsy syndrome과 연관된 약물 중 적어도 하나를

포함할 수 있다. 그리고, 단계 S210에서 추출되는 생물학적 엔티티는 레벨 별로 복수의 생물학적 엔티티를 포함할 수도 있다. 도3에서 예시한 바와 같이, 소정의 검색어가 질환명인 epilepsy syndrome인 경우, 단계 S210에서 추출되는 생물학적 엔티티는 epilepsy syndrome과 연관된 복수의 유전자, epilepsy syndrome과 연관된 복수의 단백질, epilepsy syndrome과 연관된 복수의 신진대사체, epilepsy syndrome과 연관된 복수의 증상, epilepsy syndrome과 연관된 복수의 질환, epilepsy syndrome과 연관된 복수의 화합물 및 epilepsy syndrome과 연관된 복수의 약물 중 적어도 하나를 포함할 수도 있다.

[42] 이와 같이, 단계 S210 및 단계 S220에서 DB 매트릭스를 이용하여 소정의 검색어와 연관된 생물학적 엔티티 및 상호 연관도를 추출할 경우, 탐색되어야 할 DB의 양을 현저히 줄일 수 있으며, 이에 따라 정보를 탐색하기 위한 시간 및 비용을 줄일 수 있으며, 사용자가 원하는 정보만을 추출하는 것이 가능하다.

[43] 다음으로, 데이터 생성부(130)는 단계 S210과 단계 S220에서 추출한 결과를 이용하여 제1지식 네트워크를 생성할 수 있다(S230). 도6은 일 실시예에 따라 생성된 제1지식 네트워크의 일 예이다. 원 모양은 노드를, 선은 연결선(에지)을 나타낼 수 있다. 여기서, 제1 지식 네트워크는 단계 S200에서 수신된 소정의 검색어와 단계 S210에서 추출된 생물학적 엔티티들 각각을 노드로 하며, 단계 S220에서 추출한 소정의 검색어와 생물학적 엔티티 사이의 상호 연관도 또는 생물학적 엔티티들 사이의 상호 연관도에 따라 연결선을 이용하여 복수의 노드를 연결한 그래프 형태일 수 있다. 동일한 오믹스 레벨 내 노드들이 연결선을 통해 연결될 수도 있고, 서로 다른 오믹스 레벨 내 노드들이 연결선을 통해 연결될 수 있다. 제1 지식 네트워크 내 노드 중 하나인 노드 A로부터 다른 하나인 노드 B로 가는 경로는 다양할 수 있으며, 가능한 모든 경로가 연결선에 의하여 연결될 수 있다. 여기서, 지식 네트워크는 생물학적 엔티티 간의 상호 연관도로 이루어진 네트워크로, 생물학적 네트워크로도 불릴 수 있다.

[44] 다음으로, 데이터 처리부(140)는 단계 S230에서 생성한 제1 지식 네트워크의 그래프 이론 지표를 계산할 수 있다(S240). 일 실시 예에 따라, 그래프 이론 지표는 제1 지식 네트워크를 구성하는 복수의 노드들에 대한 노드 간 최단 경로, 노드 별 클러스터링 계수, 노드 별 센트럴리티 계수, 노드 별 허브 성격 중 적어도 하나를 포함할 수 있다.

[45] 노드 간 최단 경로는 제1지식 네트워크에서 노드 A로부터 노드 B로 가는 수 많은 경로 중 가장 짧은 경로를 의미할 수 있다. 이하, 생물학적 엔티티 중 하나인 노드 A와 생물학적 엔티티 중 다른 하나인 노드 B 간 최단 경로를 산출하는 방법을 설명한다.

[46] 노드 A로부터 노드 B로 가는 경로는 다양하며, 노드 A와 노드 B가 직접 연결되거나, 노드 A와 노드 B 사이의 각 경로 상에 적어도 하나의 중간 노드가 존재할 수도 있다. 데이터 처리부(140)는 노드 A와 노드 B 사이의 최단 경로를 경로 별 중간 노드의 개수를 이용하여 획득할 수 있다. 예를 들어, 데이터

처리부(140)는, 노드 A와 노드 B 간 다양한 경로 중 중간 노드의 개수가 적을수록 짧은 경로인 것으로 판단할 수 있다.

[47] 또는, 데이터 처리부(140)는 노드 A와 노드 B 간 최단 경로는 경로 별 중간 노드의 개수를 이용하여 얻되, 연결선 별 상호 연관성의 종류를 반영할 수도 있다. 즉, 상호 연관성의 범주 별로 가중치를 다르게 설정할 수 있으며, 경로 별로 존재하는 상호 연관성에 가중치를 적용할 수도 있다.

[48] 수학식 1은 노드 간 최단 경로를 산출하는 식의 한 예이다.

[49] [수식1]

$$d_{i,j}^{W} = \sum_{w_{st} \in g_{i \to j}^{W}} f(w_{st})$$

[50] 여기서, w_{st}는 두 노드 s와 t간의 상호 연관성 지표이며, f는 가중치 변환 함수이고,

$g_{i \to j}^{W}$

는 두 노드 i와 j 사이의 최단 경로이다. 데이터 처리부(140)는 각 경로 별로 수학식 1의 값을 결정하며, 가장 낮은 값 또는 가장 높은 값을 가지는 경로를 최단 경로로서 선택할 수 있다.

[51] 다음으로, 노드 별 클러스터링 계수(clustering coefficient)는 수학식 2 및 수학식 3에 의하여 계산될 수 있다. 여기서, 클러스터링 계수는 집단화 계수라고 지칭될 수도 있으며, 특정 노드와 이웃한 노드들이 서로 연결되어 있을 확률 또는 특정 노드와 이웃한 노드들 간의 연결 밀도를 의미할 수 있다.

[52] [수식2]

$$t_i^W = \frac{1}{2} \sum_{j,h \in N} w_{ij} w_{ih} w_{jh}$$

[53] 여기서, t_i^w는 지식 네트워크의 각 노드 i 주변에 만들어지는 그래프 내의 삼각형의 개수를 의미하며, N은 지식 네트워크의 전체 노드 집합이며, w_{ij}는 노드 i와 노드 j 사이의 상호 연관성 지표이고, w_{ih}는 노드 i와 노드 h 사이의 상호 연관성 지표이며, w_{jh}는 노드 j와 노드 h 사이의 상호 연관성 지표이다.

[54] [수식3]

$$C^W = \frac{1}{n} \sum_{i \in N} \frac{2t_i^W}{k_i(k_i - 1)}$$

[55] 여기서, C^w는 클러스터링 계수를 의미하며, t_i^w는 지식 네트워크의 각 노드 i 주변에 만들어지는 그래프 내의 삼각형의 개수고, k_i는 노드 i의 degree, 즉 노드 i의 지식 네트워크 내 연결성 정도 값을 의미한다.

[56] 다음으로, 노드 별 센트렐리티(centrality) 지표는 특정 노드가 허브의 기능을 가지는지에 대한 지표이며, D_{nodal}(nodal degree)값, BC(betweenness centrality) , E_{nodal} (nodal efficiency) 값 등으로 나타낼 수 있다. 여기서, D_{nodal} 값은 각 노드의 지식 네트워크 내 연결성 정도 값, 즉, 지식 네트워크 내에서 노드 i가 얼마나 강한 또는 약한 연결성을 가지고 있는지를 나타내는 지표이고, E_{nodal} 값은 노드

i의 지식 네트워크 내 효율성 정도 값, 즉 수학식 1의 최단 경로의 역수로 표현된 값으로, 경로가 짧을수록 높은 효율성을 가지고, BC 값은 지식 네트워크 내 노드 간 경로에서 노드 i가 지름길이 되는 횟수를 나타내는 지표이다.

[57] 먼저, D_{nodal} 값은 수학식 4에 의하여 계산될 수 있다.

[58] [수식4]

$$D_{nodal}(i) = \sum_{j \in N} w_{ij}$$

[59] 여기서, wij는 노드 i와 노드 j간 상호 연관성 지표이고, N은 지식 네트워크의 전체 노드 집합이다.

[60] 그리고, E_{nodal} 값은 수학식 5에 의하여 계산될 수 있다.

[61] [수식5]

$$E_{nodal(i)} = \sum_{j \in N, j \neq i} \frac{1}{d_{i,j}^{W}}$$

[62] 여기서, N은 지식 네트워크의 전체 노드 집합이고, $d^{W}_{i,j}$는 수학식 1에서 계산한 최단 경로를 나타내는 값이다.

[63] 다음으로, Betweenness centrality(BC)는 수학식 6에 의하여 계산될 수 있다.

[64] [수식6]

$$BC(i) = \sum_{\substack{h,j \in N \\ h \neq j, h \neq i, j \neq i}} \frac{g_{hj}(i)}{g_{hj}}$$

[65] 여기서, g_{hj}는 노드 h 와 j 사이의 최단 거리를 의미하고, $g_{hj}(i)$는 노드 i를 통과하는 h 와 j 사이의 최단 거리를 의미한다.

[66] 다음으로, 소정의 노드가 허브의 기능을 가지는 것으로 판단되는 경우, 데이터 처리부(140)는 허브의 성격을 분류할 수 있다. 이때, 허브의 성격은 kinless 허브, connector 허브, provincial 허브 등으로 분류될 수 있다. 여기서, kinless 허브는 영향력이 가장 높은 허브, 즉 많은 모듈 내 노드들과 연결된 허브를 의미하고, connector 허브는 지식 네트워크 내 모듈을 연결하는 성격의 허브를 의미하며, provincial 허브는 주로 모듈 내에서 높은 영향력을 가지는 허브를 의미한다. 여기서, 모듈(module)은 전체 지식 네트워크를 세분화한 구조적 구성 그룹일 수 있다.

[67] 이를 위하여, 지식 네트워크 내의 모듈 지수(Modularity)는 수학식 7과 같이 계산될 수 있다. 모듈 지수(modularity)는 전체 지식 네트워크의 구성 모듈 종류 수를 의미한다.

[68] [수식7]

$$Q^{W} = \frac{1}{l^{W}} \sum_{i,j \in N} \left[w_{ij} - \frac{k_i^{W} k_j^{W}}{l^{W}} \right] \sigma_{mi,mj}$$

[69] 여기서,

$k_i^W = \sum_{j \in N} w_{ij}$

는 노드 i에서의 가중치 합을 의미하고,

$$l^W = \sum_{i,j \in N} w_{ij}$$

는 가중치 합을 의미한다. $\delta_{mi,mj}$ 는 크로네커의 델타(kronecker delta)이고, mi=mj인 경우 1이고, 나머지인 경우 0이다.

[70] 다음으로, 지식 네트워크 모듈의 참여지수(participation coefficient, PC)는 수학식 8과 같이 계산될 수 있다.

[71] [수식8]

$$PC_i = 1 - \sum_{m \in M} \left[\frac{k_i^W(m)}{k_i^W} \right]^2$$

[72] 여기서, M은 모듈의 집합을 의미하고,

$k_i^W(m)$

는 모듈 m 내에서 노드 i와 나머지 모든 노드 간의 연결 수를 의미하고, 모듈 m은 전체 지식 네트워크를 세분화한 구조적 구성 그룹을 의미한다.

[73] 그리고, 지식 네트워크 모듈의 z스코어(within-module degree)는 수학식 9와 같이 계산될 수 있다.

[74] [수식9]

$$z_i^W = \frac{k_i^W(m_i) - \bar{k}^W(m_i)}{\sigma_k^W(m_i)}$$

[75] 여기서, m_i 는 모듈 m 내의 노드 i를 의미하고,

$k_i^W(m_i)$

는 노드 i의 모듈 m 내에서의 연결 정도(degree)를 의미하며,

$\bar{k}^W(m_i), \sigma_k^W(m_i)$

는 각각 모듈 m내의 연결 정도 분포(degree distribution)의 평균과 표준 편차를 의미한다.

[76] 이상의 수학식 9의 지표 계산을 통해 각 노드가 모듈 내에서 허브인지 아닌지를 구분할 수 있다. 예를 들어, 다음과 같이, 지식 네트워크 모듈의 Z 스코어가 2.5 이상인 경우 허브인 것으로 판정될 수 있다.

[77] 1.within-module z-score ≥ 2.5: 허브

[78] 2. within-module z-score < 2.5: 허브 아님

[79] 또한, 노드가 모듈 내 허브인 것으로 판정될 경우, 수학식 8의 지표 계산을 통해 다음과 같이 허브의 종류를 분류할 수 있으며, 도7은 PC에 따라 허브의 종류를 구분하는 일례를 나타낸다.

[80] 1. Provincial 허브: PC ≤ 0.30

[81] 2. Connector 허브: 0.3 < PC ≤ 0.75

[82] 3. Kinless 허브: PC > 0.75

[83]

[84] 상술한 바와 같이, 데이터 처리부(140)가 단계 S240에서 그래프 이론 지표를 계산한 경우, 데이터 정제부(150)는 그래프 이론 지표를 이용하여 제1 지식 네트워크로부터 정제된 제2 지식 네트워크를 생성할 수 있다(S250).

[85] 제2 지식 네트워크는 제1 지식 네트워크보다 단순화된 네트워크로, 제1 지식 네트워크를 구성하는 복수의 노드 중 그래프 이론 측면에서 상관성이 높은 일부 노드들만으로 구성될 수 있다.

[86] 제2 지식 네트워크를 구성하는 노드는 제1 지식 네트워크를 구성하는 복수의 노드들 중 단계 S240에서 계산한 그래프 이론 지표가 기준 값 이상인 노드로 구성될 수 있다. 예를 들어, 제1지식 네트워크를 구성하는 복수의 노드들 중에서 노드 간 최단 경로에 대한 지표값, 노드 별 클러스터링 계수에 대한 지표값 및 노드 별 센트럴리티 계수에 대한 지표값 중 적어도 일부가 기준 값 이상인 일부 노드가 제2지식 네트워크에 포함될 수 있다. 즉, 제2지식 네트워크는 제1 지식 네트워크를 구성하는 복수의 노드 중에서 노드 간 최단 경로에 대한 지표값, 노드 별 클러스터링 계수에 대한 지표값 및 노드 별 센트럴리티 계수에 대한 지표값 중 적어도 일부가 임계 값 미만인 노드를 삭제하고, 삭제된 노드에 연관된 연결을 삭제하는 방법으로 생성될 수 있다.

[87] 여기서, 기준 값과 비교되는 그래프 이론 지표는 노드 간 최단 경로에 대한 지표값, 노드 별 클러스터링 계수에 대한 지표값, 노드 별 센트럴리티 계수에 대한 지표값 각각일 수 있다. 또는, 기준 값과 비교되는 그래프 이론 지표는 노드 간 최단 경로에 대한 지표값, 노드 별 클러스터링 계수에 대한 지표값, 노드 별 센트럴리티 계수에 대한 지표값 중 적어도 두 개를 통합하여 산출된 값일 수 있다.

[88] 일 실시 예에 따라, 노드 간 최단 경로에 대한 지표값, 노드 별 클러스터링 계수에 대한 지표값 및 노드 별 센트럴리티 계수에 대한 지표값 중 적어도 하나는 노드 별 표준 점수로 계산될 수 있으며, 계산된 표준 점수가 임계 값과 비교될 수 있다.

[89] 여기서, 표준 점수는 z 스코어일 수 있으며, 임계 값은 95%의 유의성을 의미할 수 있다. z 스코어는 수학식 10과 같이 계산될 수 있다.

[90] [수식10]

$$z = \frac{X - mean(x)}{SE(x)}$$

[91] 여기서, z는 z 스코어이고, X는 제1 지식 네트워크 내 특정 노드에 대한 소정의 그래프 이론 지표의 지표값이며, mean(x)는 제1 지식 네트워크 내의 적어도 일부 노드들에 대한 소정의 그래프 이론 지표의 평균 지표값이고, SE(x)는 제1 지식 네트워크 내의 적어도 일부 노드들의 그래프 이론 지표의 지표값의 표준 에러이다. 여기서, SE =

$\sigma/\sqrt{N}$

로 나타낼 수 있으며, σ는 표준 편차이고, n은 제1 지식 네트워크를 구성하는 적어도 일부 노드들의 개수이다. 일 실시 예에 따라, z 스코어를 결정하기 위해 선택되는 제1지식 네트워크의 적어도 일부 노드들의 개수는 1000개일 수 있다.

[92] 즉, z 스코어는 제1 지식 네트워크를 구성하는 각 노드에 대한 소정의 그래프 이론 지표의 지표값과 제1 지식 네트워크를 구성하는 복수의 노드에 대한 소정의 그래프 이론 지표의 평균 지표값 간의 차를 표준 에러로 나눈 값일 수 있다.

[93] 일 실시 예에 따라, z 스코어는 퍼뮤테이션 테스트(permutation test)를 통하여 계산될 수 있다. 퍼뮤테이션 테스트는 제1 지식 네트워크를 구성하는 전체 연결선을 임의로 섞은 다음, 각 노드에 대해 z 스코어를 계산하는 방법으로 행해질 수 있다. 이때, 임의로 섞는 횟수는 1000회 이상일 수 있다.

[94] 제2 지식 네트워크를 구성하는 노드는 제1 지식 네트워크를 구성하는 복수의 노드 중에서 단계 S240에서 계산한 그래프 이론 지표 중 노드 별 허브 성격에 대한 지표 값을 이용하여 추출한 일부 노드일 수도 있다. 즉, 제2 지식 네트워크를 구성하는 노드는 수학식 9의 지표 계산을 통해 모듈 내 허브인 것으로 판정된 노드, 바람직하게는 kinless 허브, connector 허브 및 provincial 허브 중 하나로 분류된 노드, 더욱 바람직하게는 kinless 허브 및 connector 허브 중 하나로 분류된 노드, 더욱 바람직하게는 kinless 허브로 분류된 노드일 수 있다.

[95] 데이터 정제부(150)는 지식 네트워크 분석 과정에서 불필요한 제1지식 네트워크의 노드를 추가적으로 제거할 수 있다. 데이터 정제부(150)는 연결선이 하나인 노드를 해당 노드의 연결선과 함께 제거할 수 있다. 연결선이 한 개에 불과한 노드는 멀티오믹스 네트워크의 개념에 부합하지 않는 네트워크 노드로 해석될 수 있기 때문이다. 또한, 데이터 정제부(150)는 클러스터링 계수(clustering coefficient)가 0인 노드를 해당 노드의 연결선과 함께 제거할 수 있다. 클러스터링 계수의 값이 0인 노드의 경우, 주요 허브 노드가 될 가능성이 없는 노드로 해석될 수 있기 때문이다.

[96] 다음으로, 출력부(160)는 단계 S250에서 생성된 제2 지식 네트워크를 출력한다(S260). 출력부(160)는, 예를 들어 디스플레이일 수 있다. 도 8은 본 발명의 실시예에 따라 "epilepsy syndrome"를 검색어로 하여 생성된 제2지식 네트워크의 일예이다. 도 8을 참조하면, 도 6의 제1 지식 네트워크에 비하여 현저히 단순화되고 정제된 제2지식 네트워크를 얻을 수 있음을 알 수 있다. 또한, 도 8을 참조하면, "epilepsy syndrome"와 연관된 서로 다른 오믹스 레벨 내 생물학적 엔티티 및 이들 간 상호 연관성을 직관적으로 얻을 수 있음을 알 수 있다.

[97] 이와 같이, 데이터 처리 장치(100)는 소정의 검색어와 관련하여 정제된 노드만으로 구성된 제2 지식 네트워크를 생성할 수 있으며, 이에 따라 신약 후보

물질 또는 신약 후보 물질의 타겟을 용이하게 결정할 수 있다.

[98] 도11은 추가적 실시예에 따른 신약 후보 물질 발굴을 위한 데이터 처리 장치의 블록도이고, 도12는 추가적 실시 예에 따른 데이터 처리 장치의 신약 후보 물질 발굴을 위한 데이터 처리 방법의 흐름도를 나타낸다.

[99] 도11 내지 12를 참조하면, 데이터 처리 장치(100)는 약물 가능 경로를 추출하기 위한 경로 추출부(180)를 더 포함할 수 있다.

[100] 여기서, 약물 가능 경로는 약물이 반응하는 경로 또는 약물이 작용하는 경로를 의미하며, 약물 반응 경로 또는 약물 작용 경로와 혼용될 수 있다. 이때, 약물 가능 경로는 서로 다른 오믹스 레벨 내 생물학적 엔티티들 간 상호 연관도에 따라 표시될 수 있으며, 본 명세서에서 생성된 제2지식 네트워크 내 일부 연결 경로를 의미할 수 있다.

[101] 경로 추출부(180)는 제2지식 네트워크에 존재하는 약물-질환 노드 페어(Pair, 쌍)들을 분석하여 신약 후보 물질을 도출하기 위한 기초 약물을 결정하기 위한 약물 가능 경로를 추출할 수 있다(S270).

[102] 도13은 일 실시 예에 따라, 데이터 처리 장치가 약물 가능 경로를 탐색하는 방법의 흐름도를 나타낸다. 도13의 흐름도는 약물 가능 경로를 추출하는 단계(S270)의 하위 단계들을 나타낼 수 있다.

[103] 단계 S13200에서, 경로 추출부(180)는 제2지식 네트워크에 존재하는 약물-질환 노드 페어들 각각에 대한 근접도의 표준 점수(z-score)가 기준 값보다 작은 약물-질환 노드 페어들을 선택할 수 있다. 경로 추출부(180)는 제2지식 네트워크로부터 특정 약물 노드과 상기 특정 약물 노드와 연결선을 통해 연결된 질환 노드를 각각 소스 노드와 타겟 노드로 하는 적어도 하나의 약물-질환 노드 페어들을 결정할 수 있다. 일 실시 예에 따라, 경로 추출부(180)는 제2지식 네트워크로부터 특정 약물에 대한 모든 약물-질환 페어들을 추출하고, 추출된 약물-질환 페어들 각각에 대한 근접도의 표준 점수를 계산할 수 있다. 일 실시 예에 따라, 노드 페어 (s, t)(s : 소스 노드(약물), t : 타겟 노드(질환))의 근접도의 표준 점수는 하기 수학식 11을 사용하여 계산될 수 있다.

[104] [수식11]

$$z(s,t)=\frac{d(s,t)-mean(d(s,T))}{SD(d(s,T))}$$

[105] (s : 소스 노드, t : 현재 타겟 노드, T : 타겟 노드들의 집합, d(s, t) : 소스 노드 s와 현재 타겟 노드 t의 최단 경로(최단 거리), mean(d(s, T)) : 소스 노드 s 와 타겟 노드 집합 T로 구성되는 노드 페어들에 대한 최단 경로들의 평균, SD(d(s, T)) : 소스 노드 s 와 타겟 노드 집합 T로 구성되는 노드 페어들에 대한 최단 경로들의 표준 편차, z(s, t) : 소스 노드 s와 현재 타겟 노드 t의 근접도의 표준 점수(z-score))

[106]

[107] 경로 추출부(180)는 근접도의 표준 점수(z-score)가 기준 값보다 작은 적어도 하나의 약물-질환 노드 페어를 선택할 수 있다. 예를 들어, 신뢰도가 90%로

설정되는 경우 기준 값은 -1.645이고, 95%로 설정되는 경우 기준 값은 -1.960이고, 99%로 설정되는 경우 기준 값은 -2.576으로 결정될 수 있다.

[108] 단계 S13400에서, 경로 추출부(180)는 단계 S13200에서 선택된 약물-질환 노드 페어의 근접도가 기준 값 이하인 페어들에 대한 경로들 중에서 경로들 각각에 존재하는 중간 노드(즉, 약물 노드와 질환 노드 사이에 존재하는 노드)가 기준 개수 이상인 경로들을 추출할 수 있다. 예를 들어, 경로 추출부(180)는 단계 S13200에서 추출된 페어들 중에서 두 개 이상의 중간 노드가 존재하는 약물-질환 노드 페어의 경로들을 추출할 수 있다.

[109] 단계 S13600에서, 경로 추출부(180)는 단계 S13400에서 추출된 중간 노드가 기준 개수 이상인 경로들 중에서 중간 노드들의 센트럴리티(centrality) 계수의 총합이 기준 값 이상인 경로를 약물 가능 경로로서 추출할 수 있다. 예를 들어, 경로 추출부(180)는 단계 S13400에서 추출된 중간 노드가 기준 개수 이상인 경로들 각각에 대해 경로를 구성하는 중간 노드들의 센트렐리티(centrality) 계수의 총합을 계산하고, 계산된 총합이 상위(예를 들어, 단계 S13400에서 추출된 경로들의 중간 노드 센트렐리티(centrality) 계수의 총합에 대한 분포 중 상위 1% 이내)인 경로들을 약물 가능 경로로서 추출할 수 있다. 이로써, 경로 추출부(180)는 제2지식 네트워크 내에서 집중도가 높은 노드를 거치고 이동 경로의 효율을 높이는 약물 가능 경로를 추출할 수 있다.

[110] 본 명세서에서 사용되는 '~부'라는 용어는 소프트웨어 또는 FPGA(field-programmable gate array) 또는 ASIC과 같은 하드웨어 구성요소를 의미하며, '~부'는 어떤 역할들을 수행한다. 그렇지만 '~부'는 소프트웨어 또는 하드웨어에 한정되는 의미는 아니다. '~부'는 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서 '~부'는 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들, 및 변수들을 포함한다. 구성요소들과 '~부'들 안에서 제공되는 기능은 더 작은 수의 구성요소들 및 '~부'들로 결합되거나 추가적인 구성요소들과 '~부'들로 더 분리될 수 있다. 뿐만 아니라, 구성요소들 및 '~부'들은 디바이스 또는 보안 멀티미디어카드 내의 하나 또는 그 이상의 CPU들을 재생시키도록 구현될 수도 있다.

[111] 한편, 상술한 데이터 처리 방법은 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의하여 읽힐 수 있는 데이터가 저장되는 모든 종류의 기록장치를 포함한다. 컴퓨터가 읽을 수 있는 기록매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플로피디스크, 광 데이터 저장장치 등을 포함할 수 있다. 또한, 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터

시스템에 분산되어, 분산방식으로 프로세서가 읽을 수 있는 코드가 저장되고 실행될 수 있다.

[112] 설명들은 본 발명을 구현하기 위한 예시적인 구성들 및 동작들을 제공하도록 의도된다. 본 발명의 기술 사상은 위에서 설명된 실시 예들뿐만 아니라, 위 실시 예들을 단순하게 변경하거나 수정하여 얻어질 수 있는 구현들도 포함할 것이다. 또한, 본 발명의 기술 사상은 위에서 설명된 실시 예들을 앞으로 용이하게 변경하거나 수정하여 달성될 수 있는 구현들도 포함할 것이다.

청구범위

[청구항 1] 데이터 처리 장치에서 수행되는 신약 후보 물질 발굴을 위한 데이터 처리 방법에 있어서,
선택된 생물학적 엔티티와 선택된 상호 연관도 종류로 구성되는 DB 매트릭스를 오믹스 DB로부터 생성하는 단계;
검색어를 수신하는 단계;
상기 DB 매트릭스로부터 상기 검색어와 다른 오믹스 레벨에 속하고 상기 검색어와 관련된 생물학적 엔티티들을 추출하는 단계;
상기 DB 매트릭스로부터 상기 검색어와 상기 생물학적 엔티티들 간의 상호 연관도를 추출하는 단계;
상기 검색어와 상기 생물학적 엔티티들 각각을 노드로 하고, 상기 검색어와 상기 생물학적 엔티티들 사이의 상호 연관도 또는 상기 생물학적 엔티티들 간 상호 연관도에 따라 연결선을 이용하여 복수의 노드들을 연결한 제1지식 네트워크를 생성하는 단계;
상기 제1지식 네트워크의 복수의 노드들 각각에 대해 그래프 이론 지표를 계산하는 단계; 및
상기 제1지식 네트워크의 복수의 노드들 중 상기 그래프 이론 지표를 이용하여 선택된 일부 노드들을 이용하여 제2지식 네트워크를 생성하는 단계를 포함하고,
상기 검색어는 유전자명, 단백질명, 신진대사체명, 증상명, 질환명, 화합물명 및 약품명 중 적어도 하나를 포함하고,
상기 생물학적 엔티티는 유전자, 단백질, 신진대사체, 증상, 질환, 화합물 및 약품 중 적어도 하나를 포함하며,
상기 상호 연관도의 범주는 참여(participate), 공변(covariate), 조절(regulate), 연관(associate), 결합(bind), 업레귤레이트(upregulate), 유사(resemble), 치료(treat), 다운레귤레이트(downregulates), 완화(palliate), 포함(include), 및 표출(express)을 포함하며,
상기 그래프 이론 지표는 상기 제1지식 네트워크를 구성하는 복수의 노드들 중 적어도 하나에 대한 노드 간 최단 경로, 노드 별 클러스터링 계수, 노드 별 센트럴리티 계수를 포함하고,
상기 연결선이 나타내는 상호연관도의 범주에 따라 상기 연결선의 가중치가 다르게 설정되고, 상기 노드 간 최단 경로는 상기 설정된 가중치를 반영하여 산출되고,
상기 제2지식 네트워크를 생성하는 단계는,
상기 제1지식 네트워크를 구성하는 복수의 노드들 각각에 대해 상기 노드 간 최단 경로, 상기 노드 별 클러스터링 계수, 및 상기 노드 별 센트럴리티 계수 중 적어도 하나에 대한 표준 점수를 계산하고, 상기 표준 점수가

임계 값 미만인 노드와 상기 임계 값 미만인 노드의 연결선을 삭제함으로써 상기 제2지식 네트워크를 생성하고,
상기 표준 점수는 제1 지식 네트워크를 구성하는 각 노드에 대한 소정의 그래프 이론 지표의 지표값과 제1 지식 네트워크를 구성하는 복수의 노드에 대한 그래프 이론 지표의 평균 지표값 간의 차를 표준 에러로 나눈 값이고,
상기 DB 매트릭스는,
상기 선택된 생물학적 엔티티들이 가로축 및 세로축 각각에 배치되며, 가로축과 세로축이 교차하는 지점에 상기 상호 연관도 종류가 표시되도록 생성되는 방법.

[청구항 2] 제1항에 있어서,
상기 제2지식 네트워크를 생성하는 단계는,
상기 제1지식 네트워크를 구성하는 전체 연결선을 임의로 섞은 다음 상기 제1지식 네트워크의 노드들 각각에 대해 상기 표준 점수를 계산하는 단계를 포함하고,
상기 임의로 섞는 회수는 1000회 이상인 방법.

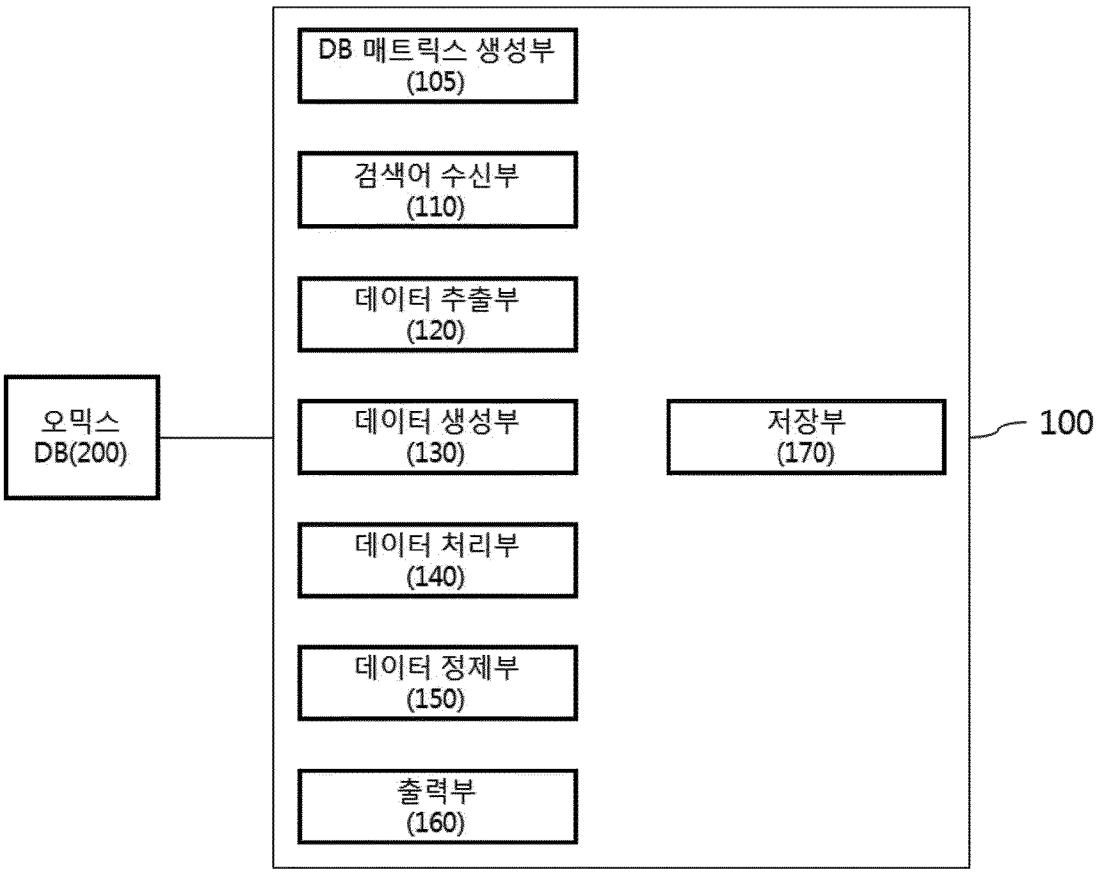
[청구항 3] 제1항에 있어서,
상기 제2지식 네트워크를 생성하는 단계는,
상기 제1지식 네트워크를 구성하는 노드들 중에서 연결선이 하나인 노드를 삭제하는 단계; 및
상기 제1지식 네트워크를 구성하는 노드들 중에서 클러스터링 계수가 0인 노드를 삭제하는 단계를 더 포함하는 방법.

[청구항 4] 제1항에 있어서,
상기 상호연관도의 범주는, 상호작용(interact), 원인(cause), 발현(present), 및 위치(localize) 중 적어도 하나를 더 포함하는 방법.

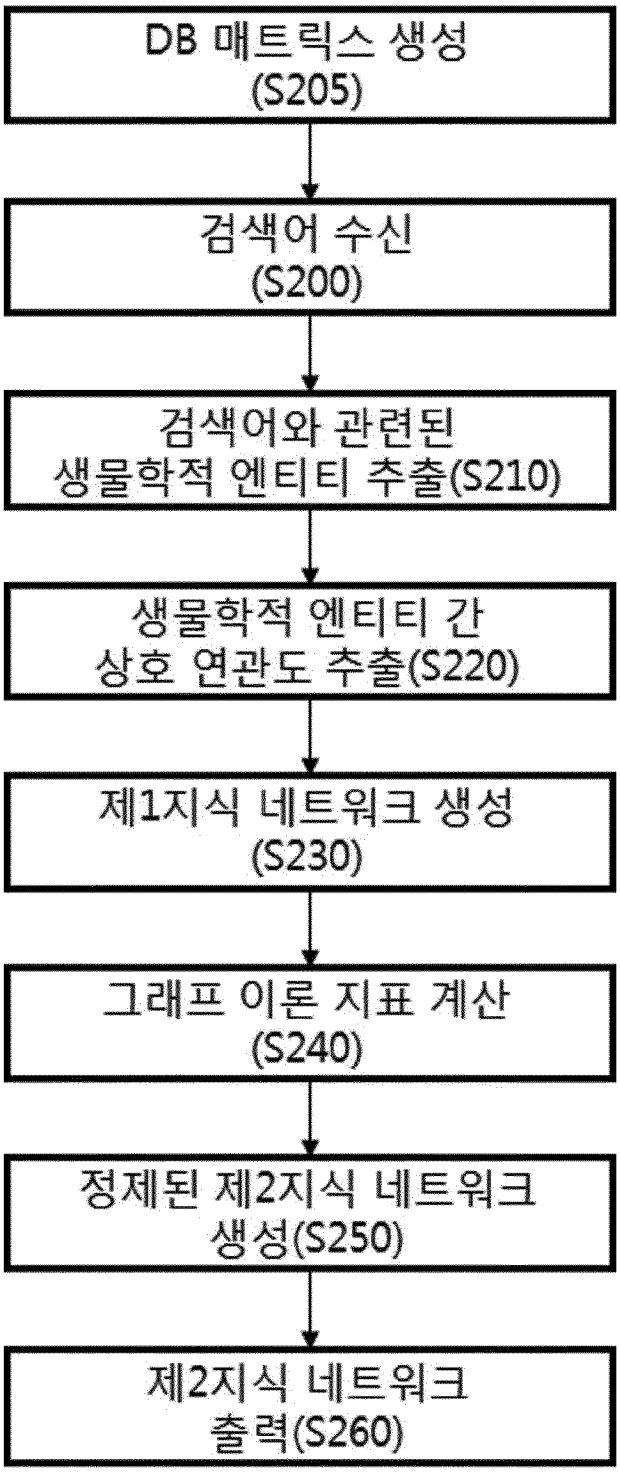
[청구항 5] 제1항에 있어서,
상기 제2지식 네트워크로부터 약물 가능 경로를 추출하는 단계를 더 포함하고,
상기 약물 가능 경로를 추출하는 단계는,
상기 제2지식 네트워크에 존재하는 약물-질환 노드들 각각에 대한 근접도의 표준 점수가 기준 값보다 작은 약물-질환 노드 페어들을 선택하는 단계;
상기 선택된 약물-질환 노드 페어들에 대한 경로들 중에서, 상기 경로들 각각에 존재하는 중간 노드가 기준 개수 이상인 경로들을 추출하는 단계; 및
상기 추출된 경로들 중에서, 상기 추출된 경로들의 중간 노드들의 센트렐리티 계수의 총합이 기준 값 이상인 경로를 상기 약물 가능 경로로서 추출하는 단계를 포함하는 방법.

[청구항 6] 제1항 내지 제5항 중 어느 한 항에서 수행되는 방법을 컴퓨터에서 실행시키기 위한 프로그램이 기록된 기록매체.

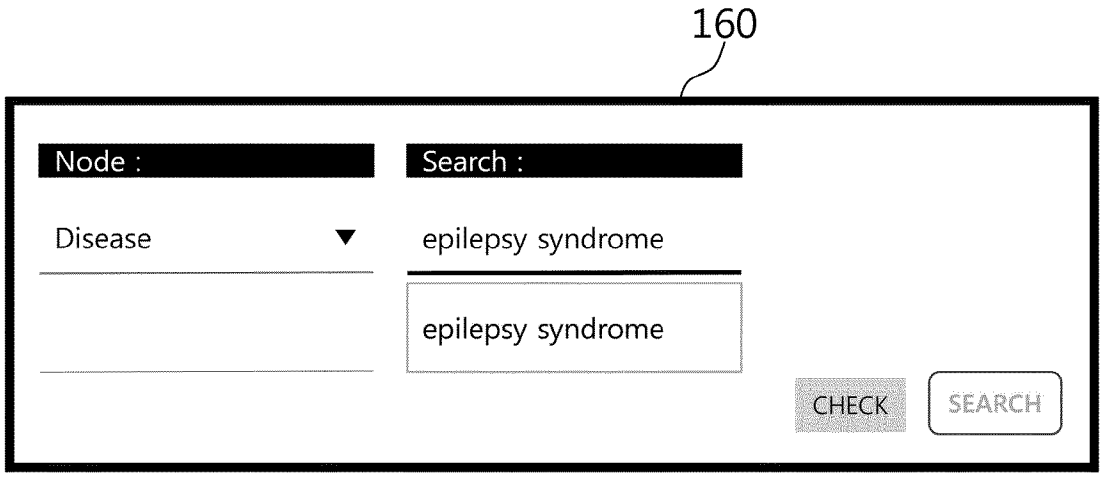
[도1]


DB 매트릭스 생성부
(105)
검색어 수신부
(110)
데이터 추출부
(120)
오믹스
DB(200)
데이터 생성부
(130)
저장부
(170)
100
데이터 처리부
(140)
데이터 정제부
(150)
출력부
(160)

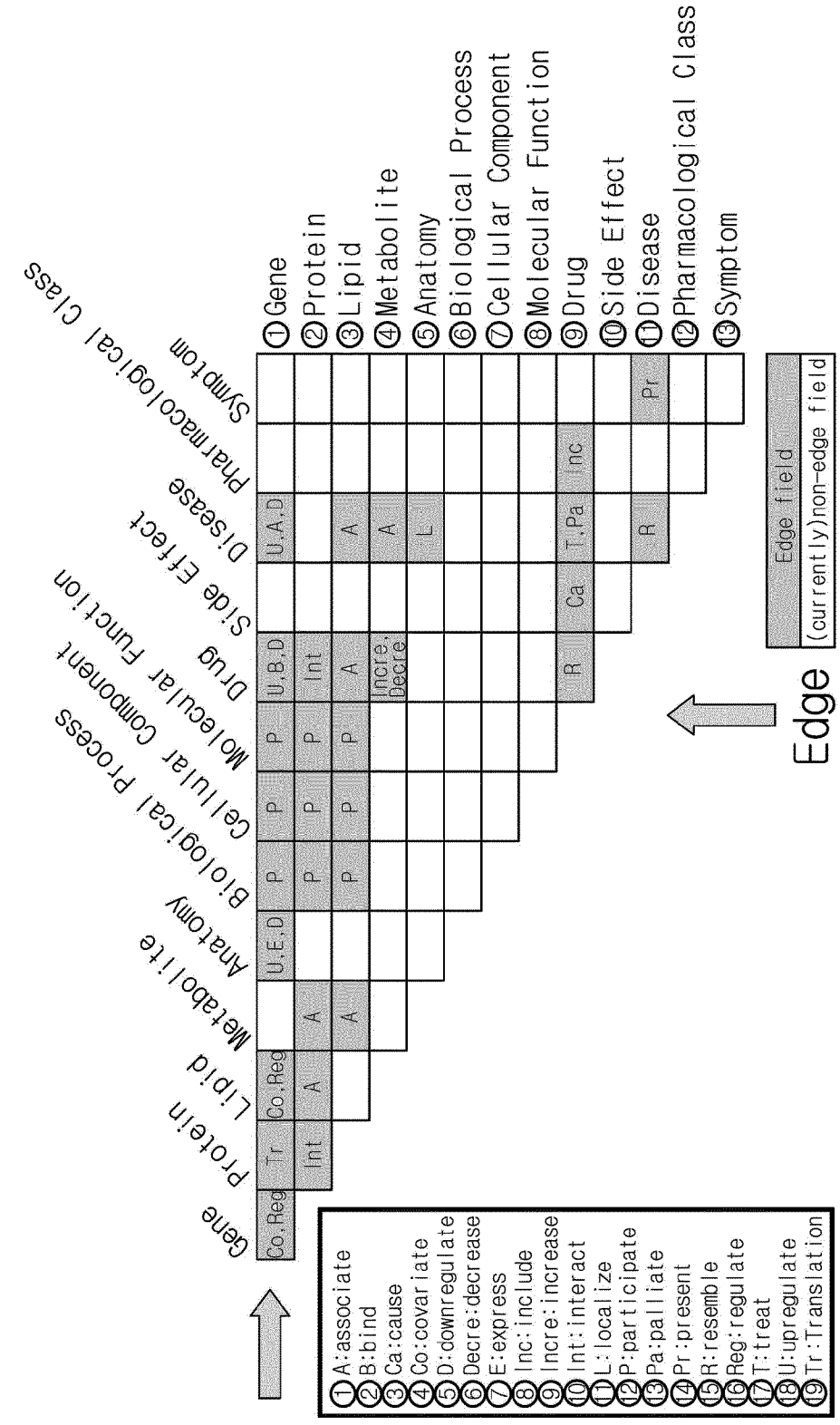
[도2]


DB 매트릭스 생성
(S205)
검색어 수신
(S200)
검색어와 관련된
생물학적 엔티티 추출(S210)
생물학적 엔티티 간
상호 연관도 추출(S220)
제1지식 네트워크 생성
(S230)
그래프 이론 지표 계산
(S240)
정제된 제2지식 네트워크
생성(S250)
제2지식 네트워크
출력(S260)

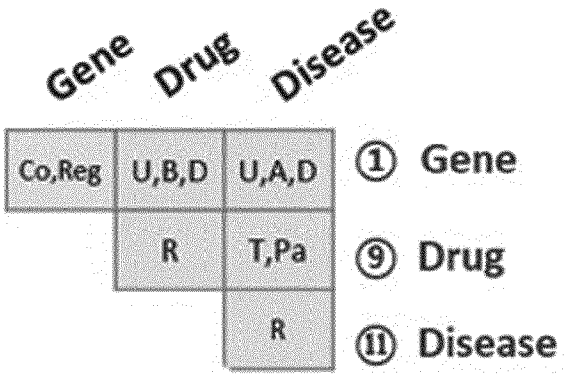
[도3]


160
Node :
Search :
Disease
epilepsy syndrome
epilepsy syndrome
CHECK
SEARCH

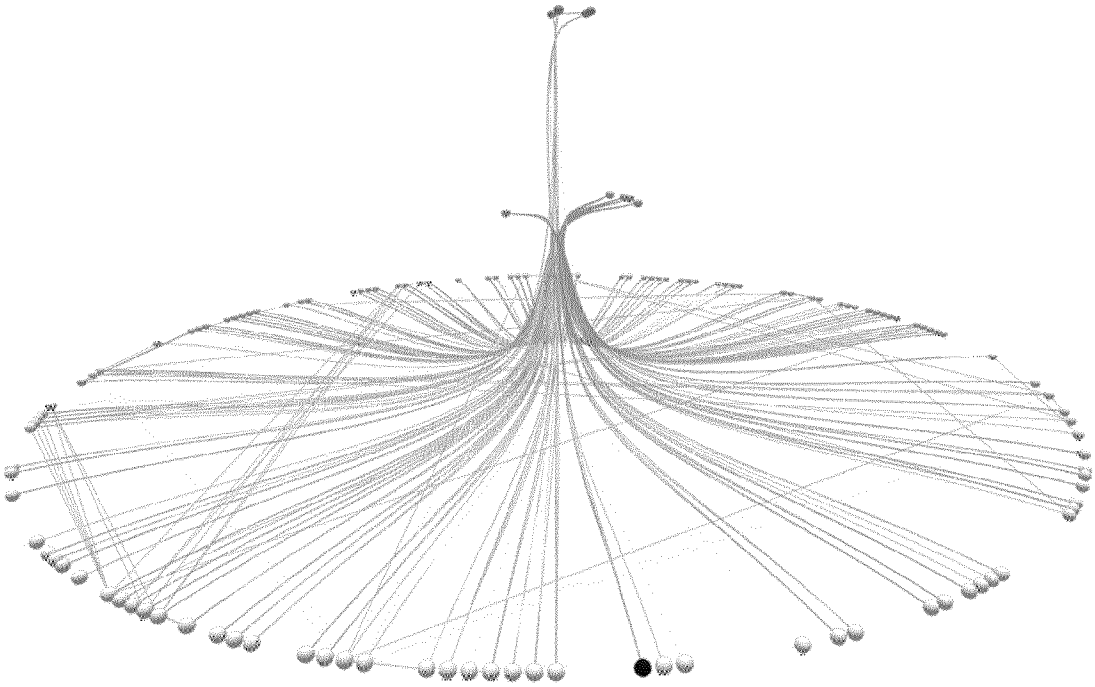
[도4]


①Gene
②Protein
③Lipid
④Metabolite
⑤Anatomy
⑥Biological Process
⑦Cellular Component
⑧Molecular Function
⑨Drug
⑩Side Effect
⑪Disease
⑫Pharmacological Class
⑬Symptom
Edge
Edge field
(currently)non-edge field
Gene
Protein
Lipid
Metabolite
Anatomy
Biological Process
Cellular Process
Molecular Component
Drug
Side Effect
Disease
Pharmacological Class
Symptom
①A:associate
②B:bind
③Ca:cause
④Co:covariate
⑤D:downregulate
⑥Decre:decrease
⑦E:express
⑧Inc:include
⑨Increase:increase
⑩Int:interact
⑪L:localize
⑫P:participate
⑬Pa:palliate
⑭Pr:present
⑮R:resemble
⑯Reg:regulate
⑰T:treat
⑱U:upregulate
⑲Tr:Translation

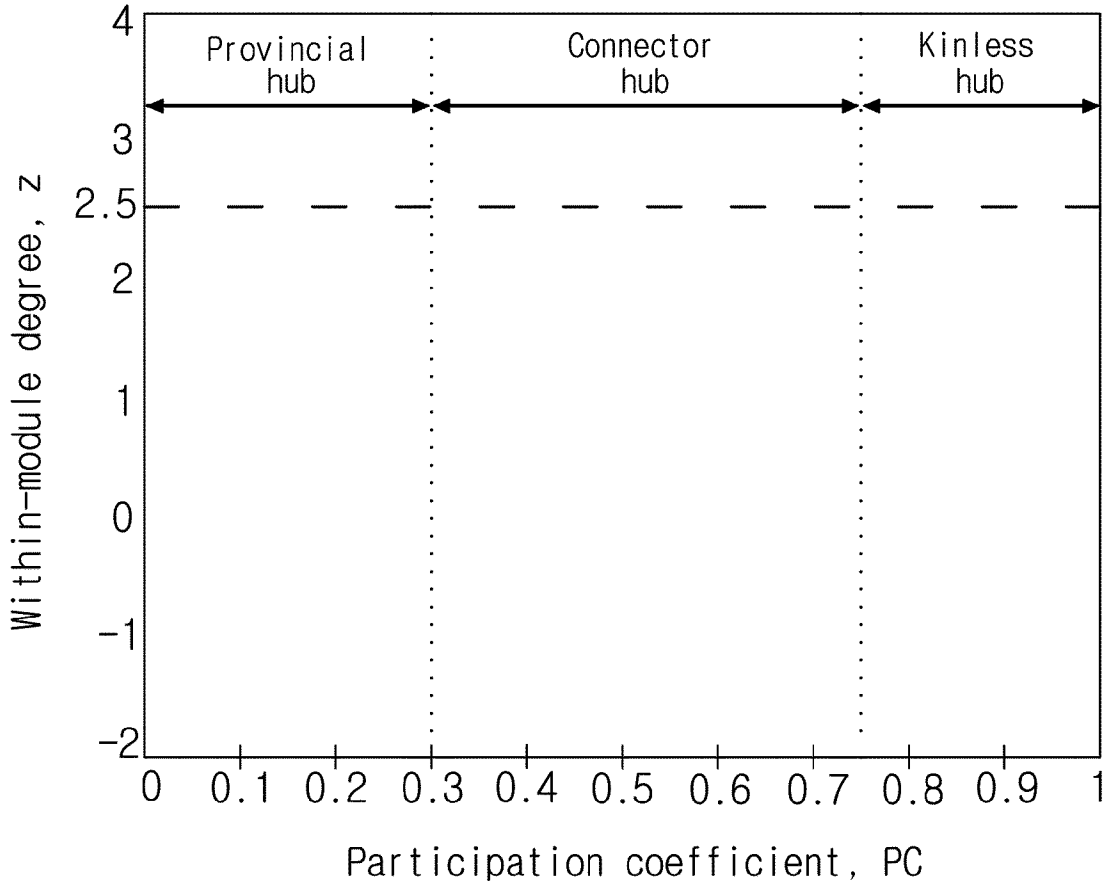
[도5]



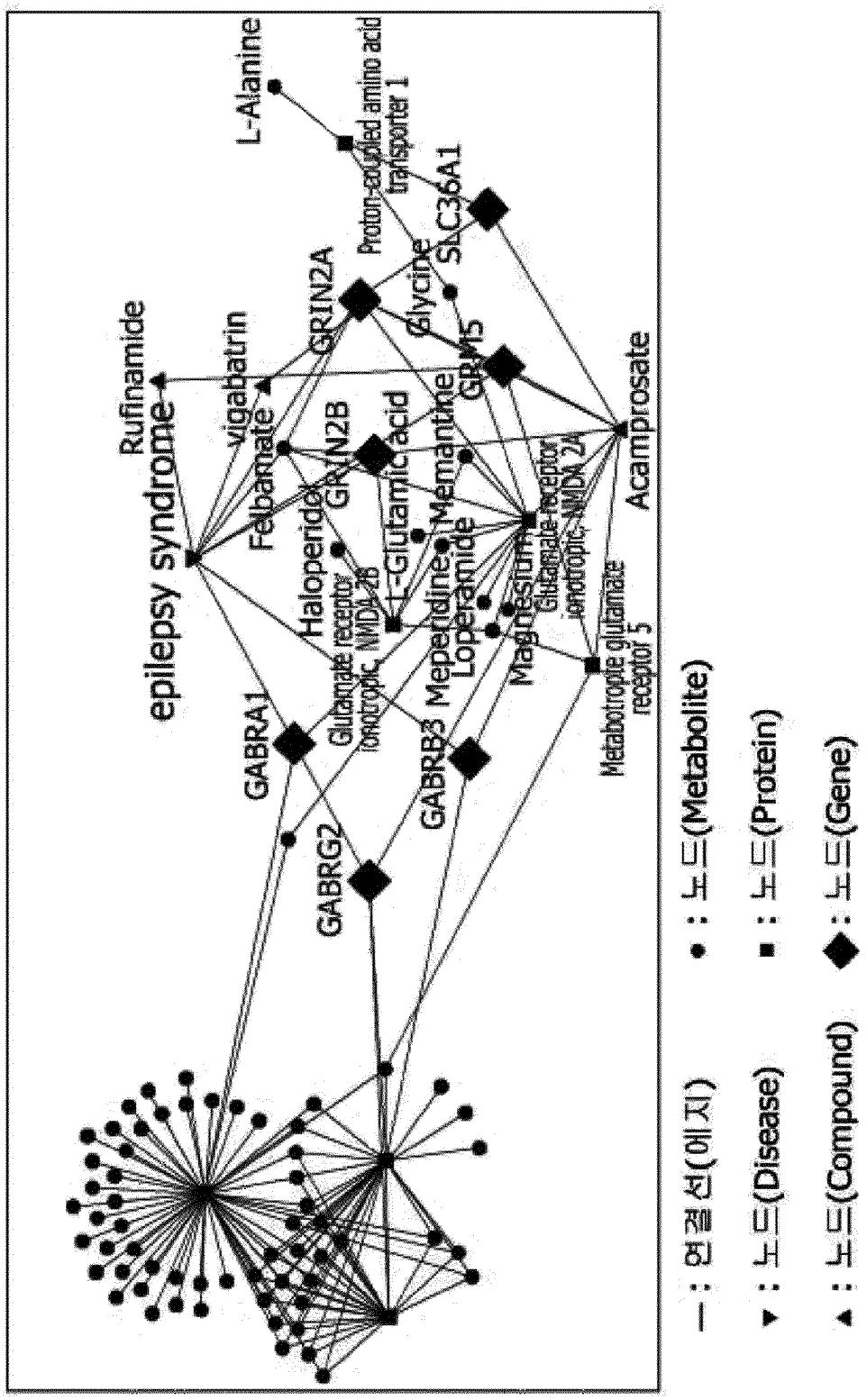
[도6]



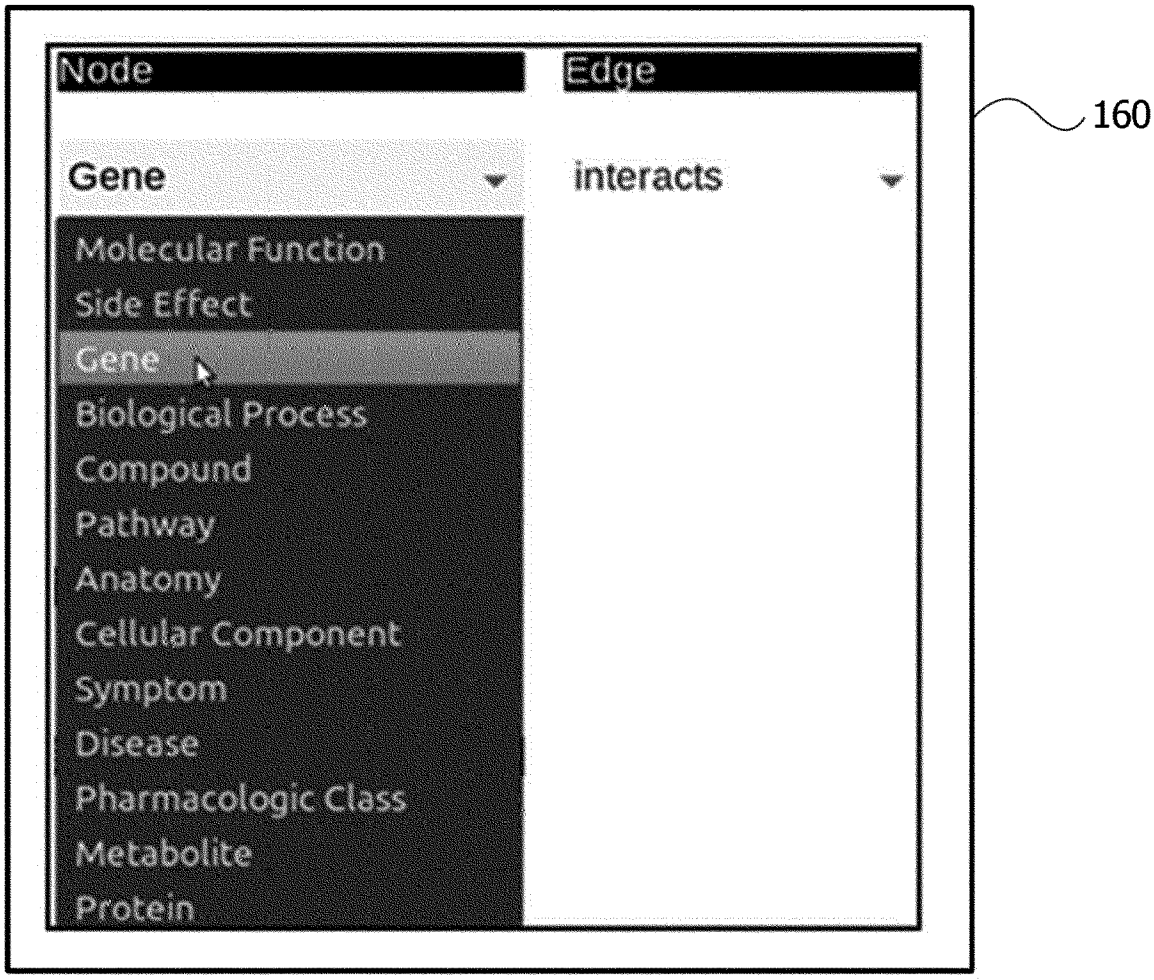
[도7]


Provincial hub
Connector hub
Kinless hub
Within-module degree, z
4
3
2.5
2
1
0
-1
-2
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Participation coefficient, PC

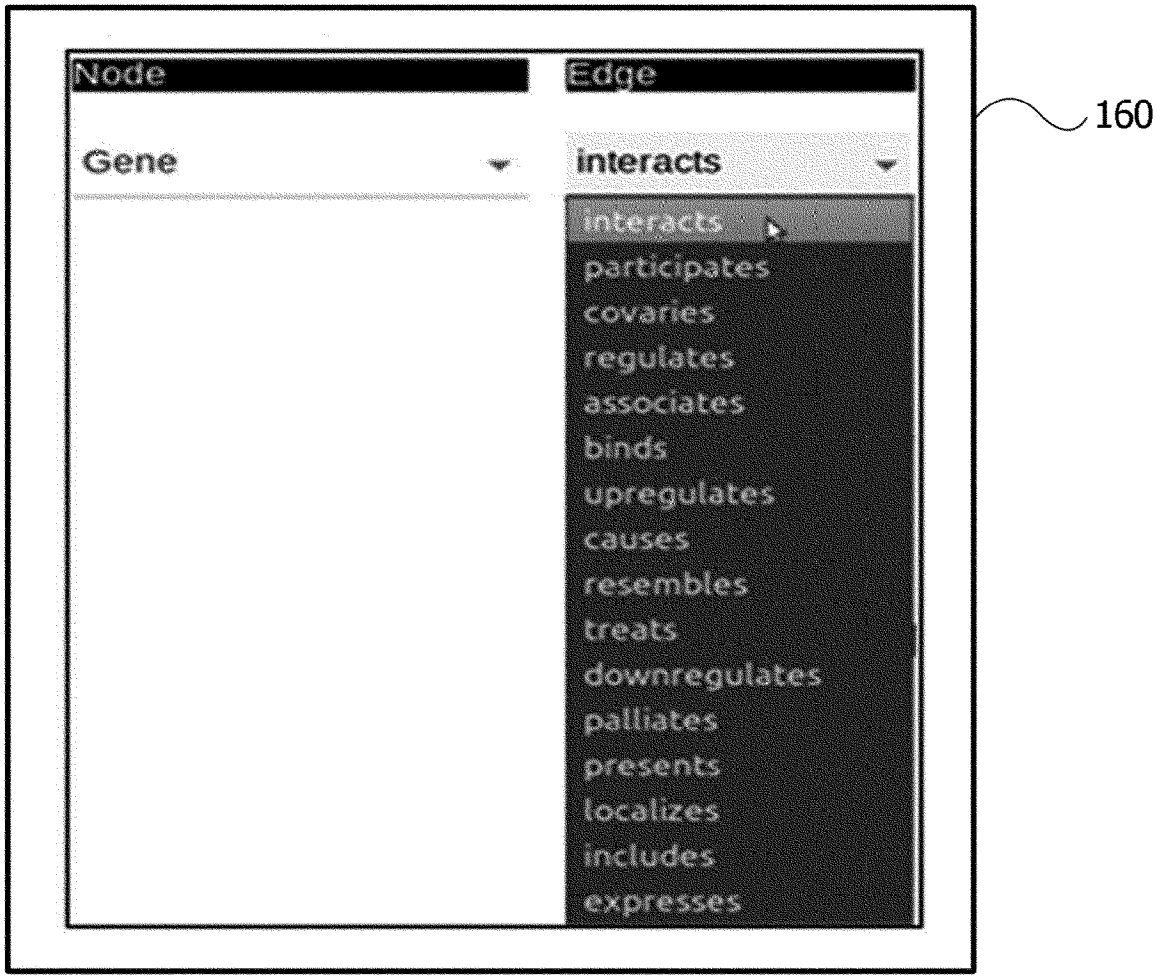
[도8]


L-Alanine
Proton-coupled amino acid transporter 1
SLC36A1
Glycine
GRIN2A
Rufinamide
vigabatrin
GRM5
Acamprosate
epilepsy syndrome
Felbamate
GRIN2B
L-Glutamic acid
Memantine
Haloperidol
Glutamate receptor ionotropic, NMDA 2B
Glutamate receptor ionotropic, NMDA 2A
Loperamide
Meperidine
Magnesium
Metabotropic glutamate receptor 5
GABRA1
GABRB3
GABRG2
—: 연결선(에지)
●: 노드(Metabolite)
■: 노드(Protein)
▼: 노드(Disease)
◆: 노드(Gene)
▲: 노드(Compound)

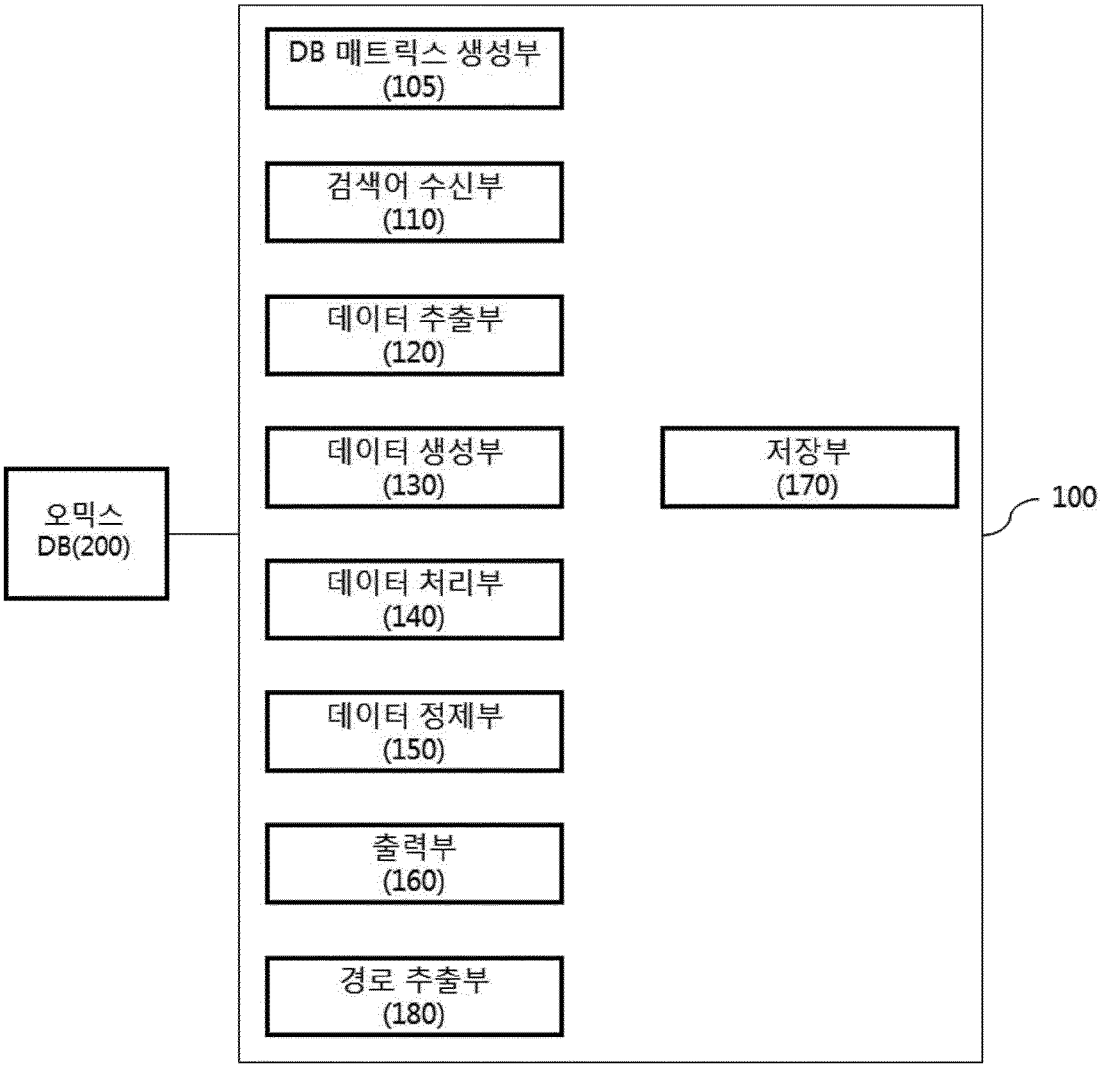
[도9]


Node
Edge
160
Gene
interacts
Molecular Function
Side Effect
Gene
Biological Process
Compound
Pathway
Anatomy
Cellular Component
Symptom
Disease
Pharmacologic Class
Metabolite
Protein

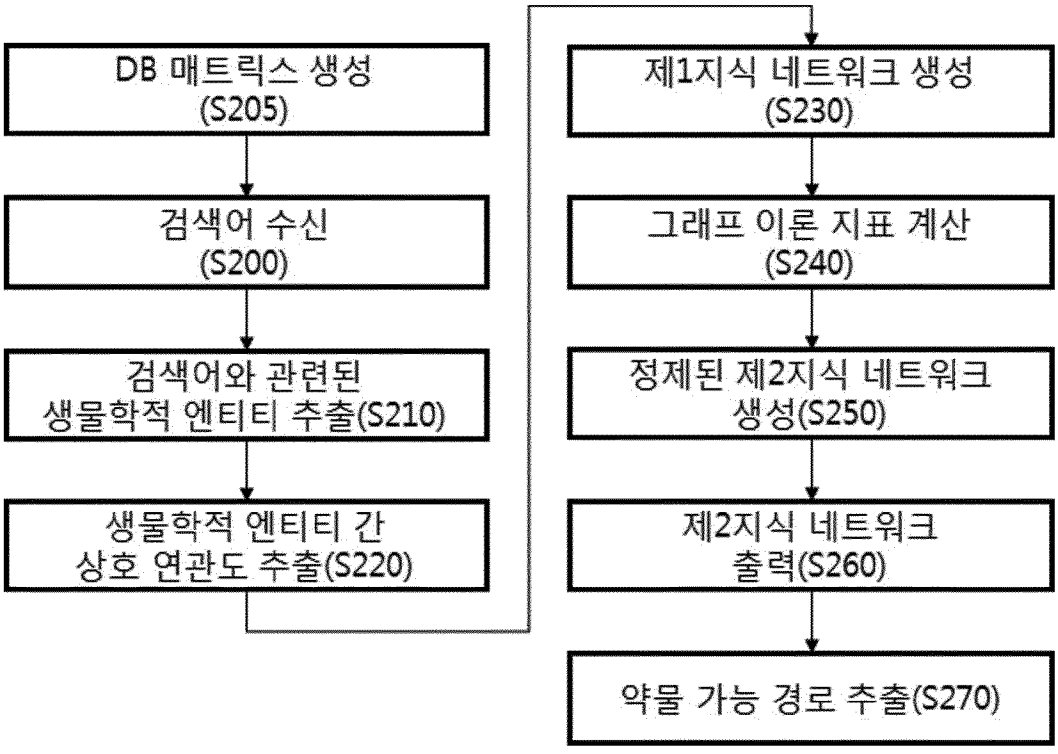
[도10]


Node
Edge
Gene
interacts
interacts
participates
covaries
regulates
associates
binds
upregulates
causes
resembles
treats
downregulates
palliates
presents
localizes
includes
expresses
160

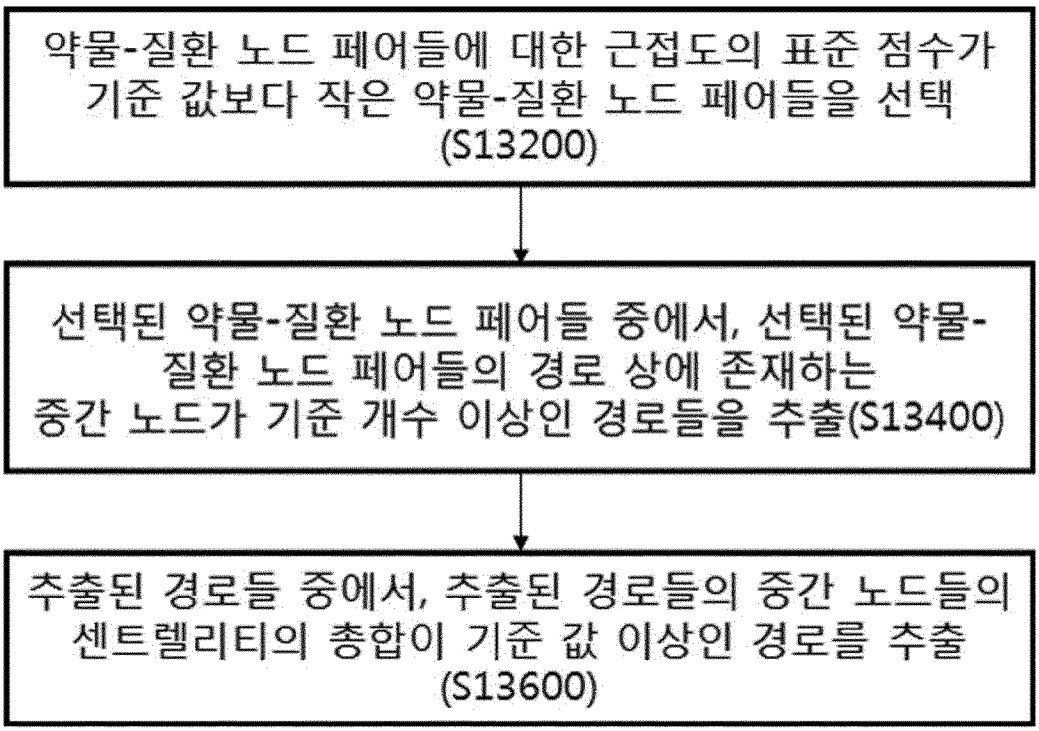
[도11]


DB 매트릭스 생성부
(105)
검색어 수신부
(110)
데이터 추출부
(120)
데이터 생성부
(130)
데이터 처리부
(140)
데이터 정제부
(150)
출력부
(160)
경로 추출부
(180)
저장부
(170)
오믹스
DB(200)
100

[도12]


DB 매트릭스 생성
(S205)
검색어 수신
(S200)
검색어와 관련된
생물학적 엔티티 추출(S210)
생물학적 엔티티 간
상호 연관도 추출(S220)
제1지식 네트워크 생성
(S230)
그래프 이론 지표 계산
(S240)
정제된 제2지식 네트워크
생성(S250)
제2지식 네트워크
출력(S260)
약물 가능 경로 추출(S270)

[도13]


약물-질환 노드 페어들에 대한 근접도의 표준 점수가 기준 값보다 작은 약물-질환 노드 페어들을 선택 (S13200)
선택된 약물-질환 노드 페어들 중에서, 선택된 약물-질환 노드 페어들의 경로 상에 존재하는 중간 노드가 기준 개수 이상인 경로들을 추출(S13400)
추출된 경로들 중에서, 추출된 경로들의 중간 노드들의 센트렐리티의 총합이 기준 값 이상인 경로를 추출 (S13600)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2019/017793

A. CLASSIFICATION OF SUBJECT MATTER

G16H 70/40(2018.01)i, G16H 50/70(2018.01)i, G16C 20/70(2019.01)i, G16C 60/00(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G16H 70/40; G01N 33/15; G06F 19/00; G16B 5/00; G16C 10/00; G16H 50/70; G16C 20/70; G16C 60/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Korean utility models and applications for utility models: IPC as above
Japanese utility models and applications for utility models: IPC as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS (KIPO internal) & Keywords: drug, omics, knowledge network, interconnectivity

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	YU, Ying et al. PreMedKB: an integrated precision medicine knowledgebase for interpreting relationships between diseases, genes, variants and drugs. Nucleic Acids Research. vol. 47, 08 November 2018(Published online), pp. D1090-D1101 <doi: 10.1093/nar/gky1042> See pages D1091-D1099; and figures 1-5.	1-6
Y	ARRELL, D. Kent et al. Network Systems Biology for Drug Discovery. Clinical Pharmacology & Therapeutics. vol. 88, no. 1, July 2010, pp. 120-125 See pages 121-124; and figure 1.	1-6
A	KR 10-1450784 B1 (AJOU UNIVERSITY INDUSTRY-ACADEMIC COOPERATION FOUNDATION) 23 October 2014 See paragraphs [0019]-[0098]; and figures 1-3.	1-6
A	KR 10-2018-0109421 A (GACHON UNIVERSITY OF INDUSTRY-ACADEMIC COOPERATION FOUNDATION) 08 October 2018 See paragraphs [0187]-[0199]; and figure 12.	1-6
A	JP 2016-099674 A (NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY) 30 May 2016 See paragraphs [0026]-[0049]; and figures 1-8.	1-6

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
31 MARCH 2020 (31.03.2020)	**31 MARCH 2020 (31.03.2020)**
Name and mailing address of the ISA/KR Korean Intellectual Property Office Government Complex Daejeon Building 4, 189, Cheongsa-ro, Seo-gu, Daejeon, 35208, Republic of Korea Facsimile No. +82-42-481-8578	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (January 2015)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/KR2019/017793

Patent document cited in search report	Publication date	Patent family member	Publication date
KR 10-1450784 B1	23/10/2014	None	
KR 10-2018-0109421 A	08/10/2018	KR 10-1964694 B1	07/08/2019
JP 2016-099674 A	30/05/2016	JP 6550571 B2	31/07/2019

국제조사보고서

국제출원번호
PCT/KR2019/017793

A. 발명이 속하는 기술분류(국제특허분류(IPC))

G16H 70/40(2018.01)i, G16H 50/70(2018.01)i, G16C 20/70(2019.01)i, G16C 60/00(2019.01)i

B. 조사된 분야

조사된 최소문헌(국제특허분류를 기재)

G16H 70/40; G01N 33/15; G06F 19/00; G16B 5/00; G16C 10/00; G16H 50/70; G16C 20/70; G16C 60/00

조사된 기술분야에 속하는 최소문헌 이외의 문헌

한국등록실용신안공보 및 한국공개실용신안공보: 조사된 최소문헌란에 기재된 IPC
일본등록실용신안공보 및 일본공개실용신안공보: 조사된 최소문헌란에 기재된 IPC

국제조사에 이용된 전산 데이터베이스(데이터베이스의 명칭 및 검색어(해당하는 경우))

eKOMPASS(특허청 내부 검색시스템) & 키워드: 약물(drug), 오믹스(omics), 지식 네트워크(knowledge network), 상호 연관도(interconnectivity)

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
Y	YING YU 등. PreMedKB: an integrated precision medicine knowledgebase for interpreting relationships between diseases, genes, variants and drugs. Nucleic Acids Research, 47권, 2018.11.08.(온라인 공개), pp. D1090-D1101 <doi: 10.1093/nar/gky1042> 페이지 D1091-D1099; 및 도면 1-5 참조.	1-6
Y	D. KENT ARRELL 등. Network Systems Biology for Drug Discovery. Clinical Pharmacology & Therapeutics, 88권, 1호, 2010.07., pp. 120-125 페이지 121-124; 및 도면 1 참조.	1-6
A	KR 10-1450784 B1 (아주대학교산학협력단) 2014.10.23 단락 [0019]-[0098]; 및 도면 1-3 참조.	1-6
A	KR 10-2018-0109421 A (가천대학교 산학협력단) 2018.10.08 단락 [0187]-[0199]; 및 도면 12 참조.	1-6
A	JP 2016-099674 A (NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY) 2016.05.30 단락 [0026]-[0049]; 및 도면 1-8 참조.	1-6

☐ 추가 문헌이 C(계속)에 기재되어 있습니다.

☒ 대응특허에 관한 별지를 참조하십시오.

* 인용된 문헌의 특별 카테고리:

"A" 특별히 관련이 없는 것으로 보이는 일반적인 기술수준을 정의한 문헌

"D" 본 국제출원에서 출원인이 인용한 문헌

"E" 국제출원일보다 빠른 출원일 또는 우선일을 가지나 국제출원일 이후에 공개된 선출원 또는 특허 문헌

"L" 우선권 주장에 의문을 제기하는 문헌 또는 다른 인용문헌의 공개일 또는 다른 특별한 이유(이유를 명시)를 밝히기 위하여 인용된 문헌

"O" 구두 개시, 사용, 전시 또는 기타 수단을 언급하고 있는 문헌

"P" 우선일 이후에 공개되었으나 국제출원일 이전에 공개된 문헌

"T" 국제출원일 또는 우선일 후에 공개된 문헌으로, 출원과 상충하지 않으며 발명의 기초가 되는 원리나 이론을 이해하기 위해 인용된 문헌

"X" 특별한 관련이 있는 문헌. 해당 문헌 하나만으로 청구된 발명의 신규성 또는 진보성이 없는 것으로 본다.

"Y" 특별한 관련이 있는 문헌. 해당 문헌이 하나 이상의 다른 문헌과 조합하는 경우로 그 조합이 당업자에게 자명한 경우 청구된 발명은 진보성이 없는 것으로 본다.

"&" 동일한 대응특허문헌에 속하는 문헌

국제조사의 실제 완료일	국제조사보고서 발송일
2020년 03월 31일 (31.03.2020)	2020년 03월 31일 (31.03.2020)

ISA/KR의 명칭 및 우편주소
대한민국 특허청
(35208) 대전광역시 서구 청사로 189, 4동 (둔산동, 정부대전청사)
팩스 번호 +82-42-481-8578

심사관
강민정
전화번호 +82-42-481-8131

국제조사보고서
대응특허에 관한 정보

국제출원번호
PCT/KR2019/017793

국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
KR 10-1450784 B1	2014/10/23	없음	
KR 10-2018-0109421 A	2018/10/08	KR 10-1964694 B1	2019/08/07
JP 2016-099674 A	2016/05/30	JP 6550571 B2	2019/07/31