치의과학석사 학위논문

생물학적 패스웨이 기반의 희귀질환 신약재창출 방법론 연구

A Study on Drug Repositioning for Rare Diseases based on Biological Pathways

2020 년 8 월

서울대학교 대학원

치의과학과 의료경영과정보학 전공

김 혜 연

생물학적 패스웨이 기반의 희귀질환 신약재창출 방법론 연구

지도교수 김 홍 기

이 논문을 치의과학석사 학위논문으로 제출함
2020 년 5 월

서울대학교 대학원
치의과학과 의료경영과정보학 전공
김 혜 연

김혜연의 치의과학석사 학위논문을 인준함
2020 년 7 월

위 원 장 류 현 모 (인)

부위원장 김 홍 기 (인)

위 원 임 정 준 (인)

국문 초록

생물학적 패스웨이 기반의
희귀질환 신약재창출 방법론 연구

김혜연
서울대학교 의료경영과정보학전공

서론: 본 연구는 생물학적 패스웨이를 활용하여 희귀질환의 신약재창출 방법론 연구를 목적으로 한다. 전 세계적으로 7,000여 개의 희귀질환이 존재하지만, 치료제는 약 5% 정도만 존재해 더 많은 연구가 필요하다. 희귀질환 치료제 연구에는 전통적인 신약개발 연구보다는 이미 승인된 약물의 새로운 의학적 용도를 찾는 신약재창출이 시간과 비용이 줄어 대안이 될 수 있다. 생물학적 패스웨이는 생체요소 간의 상호작용을 상세히 설명해둔 생물학적 심층 지식으로 유전자들의 정보를 유기적으로 바라볼 때 사용된다. 따라서 신약재창출을 위해 생물학적 패스웨이는 활용하기에 적합하다. 희귀질환의 신약재창출 약물 후보를 찾기 위해 약물 관련 유전자들의 정보와 희귀질환 관련 유전자정보를 분석하여 공통 생물학적 패스웨이 목록을 활용한다. 공통 생물학적 패스웨이로 만들어진 희귀질환과 약물의 유사도를 계산하여 희귀질환-약물 후보목록을 만든다.

방법: 희귀질환 유전자정보 데이터베이스 Panel의 희귀질환 309개와 유의미한 관련성을 가진 유전자정보를 활용하였다. 약물 데이터베이스로 DRUGBANK를 사용하였으며, 1888개의 승인된 약물과 관련된 유전자정보를 사용하였다. 패스웨이 데이터베이스로 Reactome에서 제공하는 분석 도구를 사용하여 희귀질환과 약물에 관련된 유전자 목록의 생물학적 패스웨이를 FDR 값을 기준으로 각각 수집하였다. 수집한 생물학적 패스웨이들 중 희귀질환과 약물에 공통으로 관련된 생물학적 패스웨이는 1883개로, 희귀질환과 약물의 유사도를 확인을 위해 활용된다. 희귀질환

과 약물의 유사도는 FDR값을 벡터화하여 유클리디안 유사도로 계산하였다.

결과: 본 연구 방법을 통해 희귀질환-약물 후보목록을 만들었다. 희귀질환 - 약물의 유사도 결과로 나온 값이 작은 값일수록 서로 가까운 거리에 존재한다고 설명할 수 있다. 따라서 유사도 값은 희귀질환의 신약재창출 후보가 될 가능성을 나타낸다. 이를 확인하기 위해 FDR 승인되어 희귀질환 치료제로 쓰이는 약 정보와 값을 비교하였다. “Lomitapide” 약물은 “Homozygous familial hypercholesterolemia” 질병 치료제로, 유사도 값이 2.8로 309개의 희귀질환 중 34번째로 약물-희귀질환 목록에 “Familial hypercholesterolaemia targeted panel”의 이름으로 존재했다. 희귀질환-패스웨이-약물로 분석한 결과가 희귀질환-유전자-약물의 관계보다 유의미한 결과라는 것을 “Thalidomide” 약을 통해 비교해보았다. 희귀질환-패스웨이-약물에서 “Thalidomide” 치료제가 어떤 희귀질환에 관련성이 높은지를 순서대로 볼 수 있었고, “Bladder cancer pertinent cancer susceptibility”라는 희귀질환이 가장 가까운 것으로 확인되었고, 관련 연구가 진행되었었음을 확인하였다.

결론: 희귀질환-약물 목록이 희귀질환 치료제와의 비교를 통해 연관성이 있다는 것을 확인하였고, 우선순위목록을 통해 얼마나 연관성이 있는지 알 수 있었다. 또한, 희귀질환-유전자-약물의 관계보다 희귀질환-패스웨이-약물이 더 많은 신약재창출 가능성을 가진 정보라는 것을 알 수 있었다. 따라서 희귀질환 뿐만 아니라, 다른 질병의 관련 유전자 정보를 활용하여 생물학적 패스웨이를 추출하고, 이를 신약재창출 후보를 정렬하여 알 수 있도록 기대해 볼 수 있다.

주요어: 희귀질환, 유전자, 신약재창출, 생물학적 패스웨이, 약물추천, 신약개발

학 번: 2017 - 27136

목 차

표 목 차

그림 목차

Ⅰ. 서 론

1. 연구의 필요성

희귀질환이란 유병(有病)인구가 2만 명 이하이거나 의료 진단이 어려워 유병인구를 알 수 없는 질환으로 국내 보건복지부 기준, 보건복지부령으로 정한 절차와 기준에 따라 정한 질환을 말한다[1]. 유병인구라는 것은 대상 집단에서 특정 질병의 수적 정도를 말하며, 나라마다 희귀질환에 대한 기준은 다르지만, 일반적으로 유병률이 1만 명당 5명 이하인 질병을 말한다. 희귀질환의 원인은 아직 정확하게 밝혀지지 않은 상태지만, 유전질환이 80% 정도로 유전자 또는 염색체 변화로 인한 가족력인 경우가 많다. 희귀질환은 전 세계적으로 약 7,000개가 등록되어 있으며, 매년 250개 정도의 새로운 희귀질환들이 의학저널에 보고되고 있다. 이렇듯 전 세계적으로 희귀질환을 앓는 환자들은 늘어가고 있으며, 한국 내 희귀질환 환자도 약 50만 명 정도 된다. 하지만 치료제가 존재하는 희귀질환은 약 500개 정도로 전체 희귀질환의 약 5% 정도이다. 희귀질환의 낮은 유병률로 인해 수익성이 보장되지 않아 어떤 기업도 관심을 두지 않는 분야이기 때문이다. 현재는 공공기관에서 희귀질환에 대한 연구를 적극적으로 지원해주고 있지만 [1] , 그래도 여전히 치료제는 부족한 상황이다. 희귀질병에 대한 유전자정보가 부족하고, 희귀질환 관련 임상 실험 대상자가 적어 연구개발의 어려움이 있기에 희귀질환 치료제 대한 연구를 더 적극적으로 진행되어야 할 필요가 있다. 또한, 희귀질환 연구의 어려움인 유전자정보 부족에 대한 대안이 필요하다. 본 논문은 이러한 희귀질환의 치료제의 어려움인 유전자정보 부족에 대한 대안으로 생물학적 패스웨이를 활용해 신약재창출을 하려 한다.

기존의 전통적 신약개발과정은 질병 작용점을 선정한 뒤, 약물 스크리

1) 보건복지부공고 제2018-605호 : 희귀질환관리법에 따른 희귀질환 지정

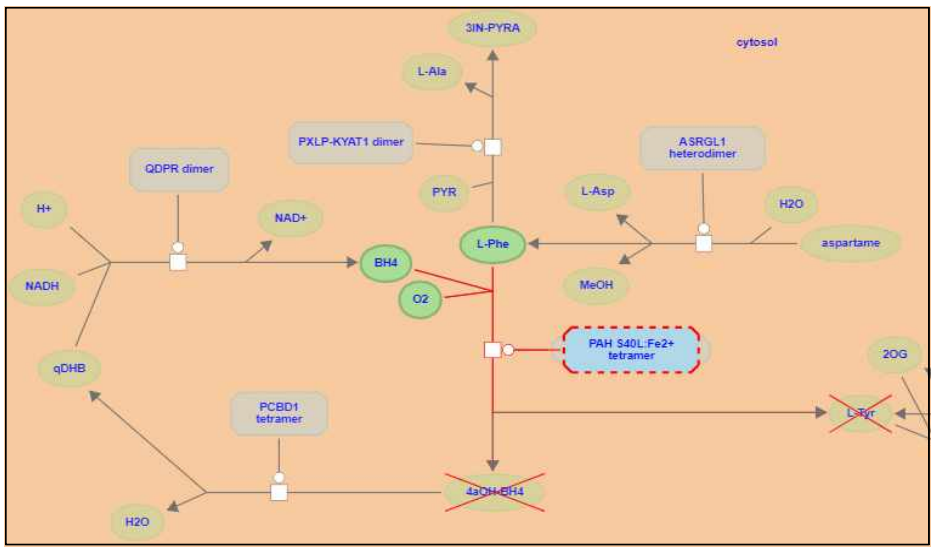
닝 후에 약물의 최적화 등의 신약 연구단계를 진행한다. 전 임상 과정, 임상 과정을 1, 2, 3단계로 진행하며 FDA 심사와 등록 등의 개발단계를 거친다. 위의 절차를 통해 총 10-15년 정도 걸리는 과정으로 평균 10억 달러 이상의 자금이 소요된다. 이렇게 많은 시간과 자본이 들어감에도 불구하고 신약개발의 가능성은 10% 내외이다.

신약재창출은 시장에서 이미 판매 중이거나 임상 단계에서 안정성 이외의 이유로 산업화에 실패한 약물들의 새로운 의학적 용도를 개발하는 신약개발의 한 방법이다. 전통적인 신약개발과정보다 기간이 절반으로 줄어들 수 있으며 이에 따라 비용 절감도 가능하다. 신약재창출은 이미 개발된 약물 중에서 선정한 약물로 개발을 시작해 임상 단계로 들어가기까지의 과정인 단순화되어 일반적으로 3-12년의 기간이 소요된다. 따라서 이미 안전성이 검증된 약물을 활용함으로써 비용 절감 및 개발 기간의 단축을 이룰 수 있다는 장점이 있다. 이러한 이유로 본 논문은 기존 치료제로 승인된 약물들에 대한 유전자 정보를 활용해 관련 생물학적 패스웨이 찾고 공통된 생물학적 패스웨이를 활용하여 희귀질환 신약재창출을 하고자 한다.

신약재창출을 하기 위해서는 두 가지의 가정이 필요하다. 첫째, 한 개의 약물이 한 개의 유전자만 조절하는 것이 아니라 다른 유전자도 조절할 수 있다는 것이다.[2] 실제로 체내에 투여하는 약물이 여러 단백질이 붙어 약효를 나타내는 경우가 많다. 예로, 아스피린은 해열과 소염진통제로도 쓰이지만, 혈전 예방약으로도 쓰인다. 둘째, 특정 유전자가 한 개의 질병에만 관여하지 않고 다양한 질병에도 관여할 수 있다.[3], [4] 세포 내에서 막 단백질, 신호전달 단백질은 한 가지 역할만 하는 것이 아니라 세포나 기관에 따라 다양한 생물학적 역할을 한다. 이러한 가정으로 기존에 신약개발은 1개의 약물, 1개의 목표, 1개의 질병의 형태로 이루어졌다면 현재는 다대다의 방식으로 신약개발을 한다. 이런 다대다 방식을 사용하기 위해서는 생물학적 패스웨이가 근거가 될 수 있다.

생물학적 패스웨이란 단백질, 유전자, 세포 등 생체요소 간의 상호작

용과 역학관계를 세밀하게 설명하는 생물학적 심층 지식을 표현한 것이다.[5] 생물학적 패스웨이는 유전자를 켜거나 끄게 할 수 있고 세포를 움직일 수 있다. 가장 일반적인 생물학적 패스웨이들은 신진대사, 유전자 발편 조절 및 신호전달과 관련이 있다. 생물학적 패스웨이는 크게 3가지로 분류되는데 대사(metabolic) 패스웨이, 유전(Genetic) 패스웨이, 신호전달(signal transduction) 패스웨이의 유형으로 분류된다. 생물학적 패스웨이에는 다양한 데이터베이스로 제공되고 있다.[6] 위의 내용을 통해 단순한 유전자정보보다 생물학적 패스웨이를 통해 다양하게 상호작용하고 있는 역학관계를 활용하는 것이 질병과 약물의 관련성을 높이는 방법이 될 수 있다고 생각하였다. 따라서 본 논문은 생물학적 패스웨이를 활용하여 희귀질환과 약물의 관계를 정리해 볼 수 있도록 노력하였다. 생물학적 패스웨이를 활용하기 위해서 FDR값을 활용하였다. 아래 [그림1]은 생물학적 패스웨이의 형태 예시이다.

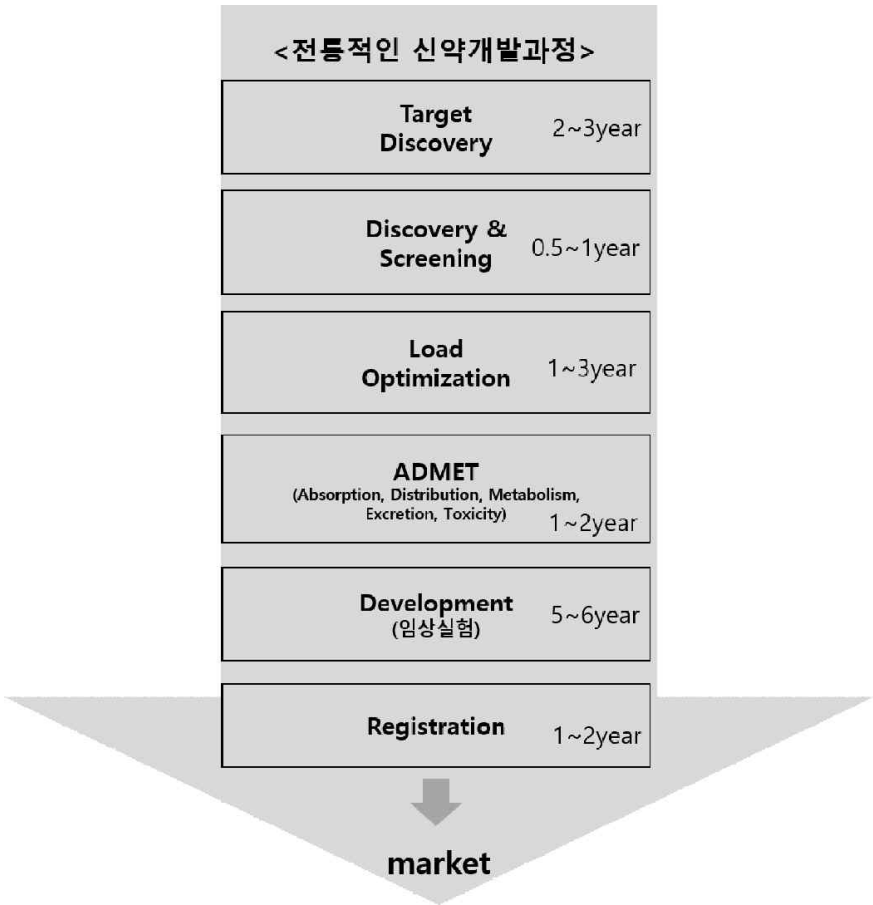


[그림1] 예시. Reactome에서 표현되는 생물학적 패스웨이의 형태

2. 신약재창출

본 장에서는 희귀질환의 신약재창출 방법론 연구로 신약재창출 방법론으로 어떠한 연구가 진행되었었는지에 대한 설명을 담았다. 전통적인 신약개발 방식과의 차이점과 현재 신약재창출 계산적인 방법에 대한 문헌 내용을 간략히 정리하였다.

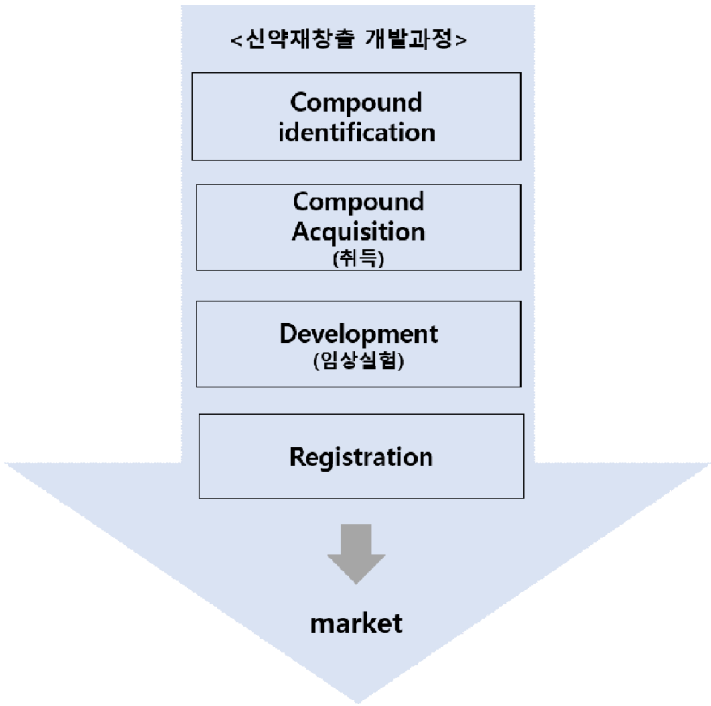
전통적인 신약개발 방식은 아래의 그림과 같다.[7]



[그림2] 전통적인 신약개발 과정

전통적인 신약개발과정은 질병 작용점을 선정한 뒤, 약물 스크리닝 후에 약물의 최적화 등의 신약 연구단계를 진행한다. Target Discovery는 임상적으로 검증된 표적에 효과를 보이는 화합물을 발견하는 과정을 설명한다. Discovery & Screening과 Load Optimization을 통해서 약 1,000개의 화합물 후보군에서 5개 정도가 전 임상 단계로 간다. ADMET(Absorption, Distribution, Metabolism, Excretion, Toxicity)은 전 임상 과정으로 실험실 및 동물 실험을 통해 진행된다. Development는 사람 대상으로 임상 시험 1상, 2상, 3상을 진행한다. Registration은 FDA 심사와 등록을 말한다. 위의 절차를 통해 총 10-15년 정도 걸리는 과정으로 평균 10억 달러 이상의 자금이 소요된다. 이렇게 많은 시간과 자본이 들어감에도 불구하고 신약개발의 가능성은 10% 내외이다.

신약재창출 방식은 전통적인 방식보다 간편한 구조로 되어있다. [8] 신약재창출의 과정은 아래의 그림과 같다. Compound Identification은 이미 시판 중이 약이나 임상 단계에서 안정성 이외의 이유로 시판되지 못했던 실패한 약물들의 가지고 후보군을 선정한다. Compound Acquisition을 통해 목표로 하는 후보군에 대한 안정성 등에 대한 정보 취득한 후, 전통적인 신약 방식과 동일하게 진행된다. 따라서 전임상과정까지의 과정을 단축해 시간과 비용을 절약할 수 있게 된다. 제약회사들은 이러한 이유로 신약재창출의 개발과정을 좀 더 활용하려는 추세이다. 신약재창출의 가장 큰 장점은 안전성이 검증된 약물을 활용하려고 하는 것이 신약재창출에서 비용과 개발 기간을 단축한다는 점이다. 전통적인 신약개발과정 중에서 안정성의 검사까지는 아무 문제가 없었지만, 약으로 승인되지 못했던 약들을 활용하는 방식과 이미 약으로 활용되고 있지만, 목적으로 했던 질병이 아닌 다른 질병에도 효과가 있을 수 있는 약에 대해 개발을 하는 과정으로 신약재창출은 진행된다. 전통적인 신약재창출의 방식보다 최소 5-7년까지 단축 가능해 더욱 유용하다. 따라서 신약재창출은 기존에 시판된 약물이나, 안정성 이외의 이유로 시판되지 못한 약물의 새로운 의학적 용도를 개발하는 과정이다.



[그림3] 신약재창출 개발과정

신약재창출에는 다양한 방식이 존재한다. 그중에서도 신약재창출을 계산적인 접근 방식일 때는 3가지로 분류된다. 네트워크 방식과 자연어 처리방식과 시맨틱 처리방식, 머신러닝 방식이다.

네트워크 기반의 방식은 여러 데이터를 통합해 신약재창출에 사용되는 방식으로 네트워크 클러스터 접근 방식과 네트워크 전파방식으로 분류된다. 네트워크 클러스터 접근 방식은 약물-질병, 약물-약물, 약물-단백질 관계와 같은 다양한 생물학적 개체의 유사한 특성을 공유한다는 사실을 기반으로 접근한 방식이다. 새로운 약물-질병 관계나 약물-단백질 관계를 발견하기 위해 사용된 방식으로 유용하게 사용된다. 네트워크 전파 접근 방식은 이전 정보의 정보가 노드로 존재하여, 노드가 전체 또는 일부의 네트워크에 전달되는 방식을 말합니다. 여러 연구에서 이렇게 전파 접근법을 사용하였을 때 질병-약물 관계를 찾는 것에 효과적이라는

것을 알 수 있었다.〔9〕

자연어처리 접근 방식은 의학, 생물학적 문헌에 대한 문헌들을 통해서 생물학적 개체 관계를 추출하기 위해 개발되고 있는 방식이다.〔10〕 대표적인 방법은 IBM에서 개발한 왓슨이 있다. 계속 새롭게 나오고 있는 의학, 생물학적 문헌들을 통해서 유전자와 질병의 관계정보를 정립하기 위한 다양한 가능성들을 확인할 수 있는 방식이다.

시맨틱 기반 접근 방식은 정보나 이미지 검색 등에 사용되는 방식이다. 대규모 의료데이터베이스의 정보들을 통해서 생물학적 개체 관계를 추출하고, 기존의 온톨로지 네트워크와 합쳐 구성된다. 이렇게 구성하기 위해서는 대규모의 의료데이터베이스가 구축되어 있어야 하며, 그 의료데이터베이스에서 필요한 정보만을 통해 활용해야 하기에 처음에 시도하기에는 어려움이 있다. 하지만 정보만 확실하게 존재하면 유사한 약물이 유사한 단백질과 상관관계가 있다는 것을 시맨틱 네트워크를 통해서 구성할 수 있다. 〔11〕

머신러닝 기반의 접근 방식은 선형회귀, SVM, 랜덤 포레스트, 뉴럴네트워크, 딥러닝 등의 방식을 말하며 요즘 최신기술로 많이 사용되는 기술이다.〔12〕 머신러닝의 기반으로 활용하기 위해서 어떤 결과를 내고 싶은지에 대한 확실한 목적에 따라 데이터 수집을 해야 한다. 데이터에 의존도가 높기에, 데이터에 대한 신뢰가 바탕으로 이루어져야 한다. 실험데이터 자체는 통일된 단위의 값을 갖지 않는 경우가 많지 않아 바로 활용할 수는 없다. 따라서 데이터 전처리과정을 통해 목적에 맞게 데이터 정보를 정제화하는 작업이 필요하다.

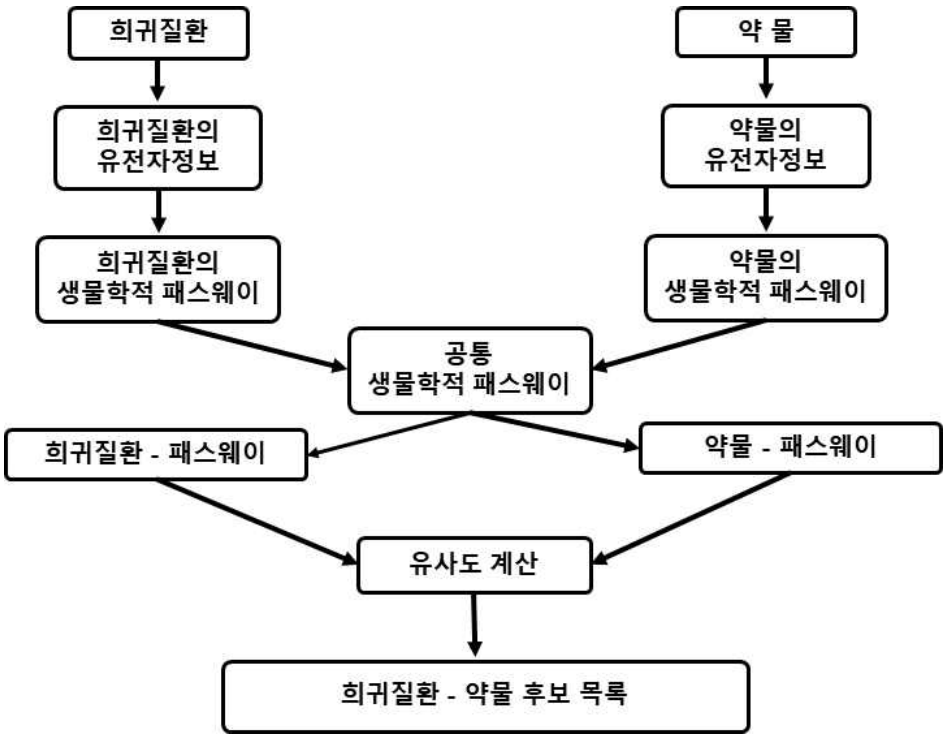
신약재창출을 하는 전략은 6가지로 분류된다. 지식 기반의 신약재창출, 표적 기반의 신약재창출, 패스웨이 기반의 신약 재창출, 표적 기술 기반의 신약재창출, 시그니처 기반의 신약재창출, 페노타입 기반의 신약재창출이다.〔13〕

이 중에서 본 논문에서도 진행하는 패스웨이 기반의 신약재창출을 한 연구만 살펴보기로 한다. 패스웨이 기반으로 신약재창출을 시도하는 다

양한 방식이 존재한다. 오믹스 데이터와 약물의 페노타입을 활용하여 신약재창출을 한 연구 [14], 유방암에 관련 특전 유전자들의 패스웨이를 살펴본 연구 [15], 실험 셀에 대한 유전자표현정보를 패스웨이로 가져오는 방식 [16] 등 다양하게 패스웨이를 활용하여 신약재창출을 위한 시도는 계속되고 있다.

3. 연구 목적

희귀질환 신약재창출을 위한 생물학적 패스웨이를 활용한 새로운 방법론을 제안한다. 아래 [그림4] 모식도 형태로 연구가 진행된다. 전처리 과정으로 희귀질환과 약물에 대한 유전자정보를 활용해 생물학적 패스웨이를 찾는다. 그 후, 공통 생물학적 패스웨이를 FDR값을 기준으로 희귀질환과 약물의 각각의 테이블을 만든다. 각각으로 만들어진 테이블은 벡터로 치환되어, 공통 패스웨이를 기준으로 유사도를 FDR값을 기준으로 계산한다. 유사도 계산을 통해서 희귀질환-약물의 후보목록을 얻을 수 있다. 따라서 결과를 통해서 유사도를 확인할 수 있어 신약재창출의 연구 방법론으로 제안한다.



[그림4] 전체 모식도

Ⅱ. 연구 재료 및 방법

1. 연구 재료

본 장에서는 논문에서 활용하는 데이터베이스들을 설명하고, 해당 데이터베이스들을 사용하는 이유와 관련 데이터들이 어떻게 사용되는지를 설명한다.

1) 희귀질환 데이터베이스

PanelApp(panel)은 희귀질환 관련 유전자 데이터베이스로, 어떤 유전자가 희귀질병에 관한 증거가 될 수 있는지에 대해 전 세계 과학계의 전문가들이 유전자 및 게놈 개체를 추가하거나 검토해 정보를 올려둔 곳으로 공개적으로 이용 가능한 자료이다.[17] panel에서는 3가지로 등급을 나누어서 질병에 관한 증거가 될 수 있는 유전자들을 분류해두었다. 3가지 등급은 녹색, 노란색, 빨간색으로 구별된다. 그중 녹색 등급은, 유전자와 질병 간의 관계에 가장 높은 증거를 갖고 있으며 해당 유전자가 게놈 해석을 위해 사용될 수 있는 유전자들을 말한다. 정리하면, 녹색 등급은 해당 질병에 가장 높은 증거를 가진 유전자들을 설명한다. 녹색 등급에 관한 가이드라인은 해당 홈페이지에서 제공되어 있으며[2] 이러한 상세한 조건들로 해당 유전자에 대한 신뢰도를 만든다. 노란색은 아직 밝혀지진 않았지만, 가능성이 있거나 애매한 유전자들을 말한다. 빨간색으로 된 유전자들은 관련성이 아직 밝혀진 바가 없다. 따라서 빨간색으로 표시된 유전자들은 정보가 정확하지 않기에, 개수가 매우 적은 편이다. 본 연구

2) panel 사이트에 있는 가이드라인
https://panelapp.genomicsengland.co.uk/#!Guidelines

는 가장 관련성이 높은 유전자들을 활용하여 더 정확한 생물학적 패스웨이를 얻을 수 있다고 생각하였기에 녹색 등급을 활용한다. 희귀질환에 관한 데이터베이스는 여러 가지가 존재하지만, panel에 있는 희귀질환을 사용한 이유는 가장 관련성 높은 유전자에 대한 정보만을 이용해 유전자 데이터에 대한 신용을 높이고자 하였다. 희귀질환에 대한 정보는 환자 수가 적고, 밝혀지지 않은 내용이 많아 유전자에 대한 명확한 정보를 얻기 어렵다. 따라서 panel에 있는 희귀질환에 대한 유전자들이 가장 활용도가 높은 데이터라고 생각하였다.

본 연구에서는 희귀질환에 유의미한 유전자들을 사용하기 위해서 panel에 있는 희귀질환 308개를 사용하였으며, 생물학적 패스웨이와 연관을 지었을 때 301개의 희귀질환을 이용하였다. 309개 중에서 사용하지 못한 8개의 희귀질환은 아직 밝혀진 유전자가 없어 희귀질환 이름만 있는 경우와 생물학적 패스웨이에서 찾을 수 없는 유전자였다. 본 연구의 데이터는 2020년 3월 17일 기준으로 panel 웹사이트에서 제공된 데이터베이스를 활용하였다.

2) 약물 데이터베이스

Drug bank는 2006년부터 지속적인 업데이트가 되는 약물 데이터베이스이다.[18] 현재(버전5.1.6, released 2020/04/22)까지 공개되어 있으며, 캐나다 건강 연구 기관(Canadian institutes of Health Research)의 지원을 받아 앨버타대학교의 The Metabolomics Innovation Centre(TMIC)에서 여전히 연구가 진행되고 있다. 현재 Drug bank 데이터베이스에 있는 소분자약물(Small Molecule Drug)의 개수는 11,388개이고, 바이오테크놀로지약물(Biotech Drug)의 개수는 2,155개이다. 데이터베이스에 있는 전체 약물 중에서 승인된 약물은 4,002개이다. 본 연구에서는 DrugBank에서 제공하는 승인된 약물과 표적 유전자들을 이용한다. DrugBank는 약

물에 대한 정보를 약물 데이터베이스 중에서 가장 많이 가지고 있으며, 업데이트도 꾸준히 유지되고 있어 데이터의 신뢰도를 위해 해당 데이터베이스를 사용하였다. DrugBank에서 승인된 약물 2134개에 대해 2782개의 유전자가 10984개로 pair를 이루고 있다. 2134개 중에서 관련있는 생물학적 패스웨이가 있는 약물은 1888개였다. 2134개에서 1888개로 예외가 된 약물들은, 관련 있는 패스웨이에서 아직 나타나지 않은 약물들로 본 연구에서는 사용할 수 없었다. 본 연구의 데이터는 2020년 4월 23일 기준으로 DrugBank에서 제공된 데이터를 사용하였다.

	DRUGBANK_약물	유전자
유일 값	2134	2782

[표1] DRUGBANK의 유전자 유일 값

3) 생물학적 패스웨이 데이터베이스

Reatome은 신호 및 대사 분자 및 생물학적 경로와 과정으로 구성된 관계에 대한 무료로 제공되는 오픈소스 관계형 데이터베이스이다.[19] Reatome은 웹사이트에서 분석서비스를 제공하고 있으며 꾸준히 업데이트된다. 알고 싶은 유전자들을 해당 웹사이트의 분석서비스에 넣으면 해당 유전자와 관련 있는 생물학적 패스웨이를 제공한다. 해당 유전자가 관련 패스웨이와 얼마나 관련이 있는지에 대해 p-value, FDR 두 가지의 방법으로 정렬하여 볼 수 있다. 상호작용하는 것에 대해 알아볼 수 있으며, 사람이 아닌 다른 종에 대해 검색이 가능하다. Reatome 데이터베이스를 사용하는 이유는 생물학적 패스웨이의 반응에 참여하는 유전자들에

대한 연산 프로세스가 잘되어있기 때문이다. 또한, API의 제공으로 분석 정보에 대한 획득도 빠르게 가능하다.

본 논문에서는 Reactome에서 서비스로 제공하는 restfulAPI를 활용해 관련 유전자들에 대한 생물학적 패스웨이를 찾아보았다. restfulAPI는 파이썬을 활용해 프로그래밍하였고, 유전자 목록은 전처리과정을 통해서 만들어 둔 것을 활용했다. 조건은 FDR 값을 기준으로, 사람을 기준으로, 인터랙션(interaction)을 False로 하였다. 총 301개의 질병과 2134개의 약물에 대해서 생물학적 패스웨이 정보를 추출하였다. [표2]는 희귀질환 309개에 대한 생물학적 패스웨이 2250개의 유일 값을 보여준다. [표3]은 약물 1888개에 대한 생물학적 패스웨이 1935개의 유일 값에 대한 설명이다. [그림5]는 RestfulAPI에 대한 예시 코드 설명이다

.

	희귀질환	생물학적 패스웨이
유일 값	309	2250

[표2] 희귀질환-생물학적 패스웨이 유일 값

	DRUGBANK_약물	생물학적 패스웨이
유일 값	1888	1935

[표3] 약물-생물학적 패스웨이 유일 값

Reactome에서 제공되는 Restful API

```
curl -X POST
"https://reactome.org/AnalysisService/identifiers/?interactors=false&species=Homo%20sapiens&pageSize=20
&page=1&sortBy=ENTITIES_FDR&order=ASC&resource=TOTAL&pValue=1&includeDisease=true"
-H "accept: application/json" -H "content-type: text/plain" -d "ATP1A3, DFNB59, OPA1, OTOF, DIAPH3"
```

Python 코드로 변환 시 작성되는 코드

Test 예시 : ATP1A3, DFNB59, OPA1, OTOF, DIAPH3

In [5]:

```python
import requests

headers = {
    'accept': 'application/json',
    'content-type': 'text/plain',
}

params = (
    ('interactors', 'false'),
    ('species', 'homo sapiens'),
    ('pageSize', '3000'),
    ('page', '1'),
    ('sortBy', 'ENTITIES_FDR'),
    ('order', 'ASC'),
    ('resource', 'TOTAL'),
    ('pValue', '1'),
    ('includeDisease', 'true'),
)

data = 'ATP1A3, DFNB59, OPA1, OTOF, DIAPH3'

response = requests.post('https://reactome.org/AnalysisService/identifiers/', headers=headers, params=params, data=data)

json_data= response.json()
```

[그림5] Restful API 활용 시 작성되는 예시 코드

2. 방 법

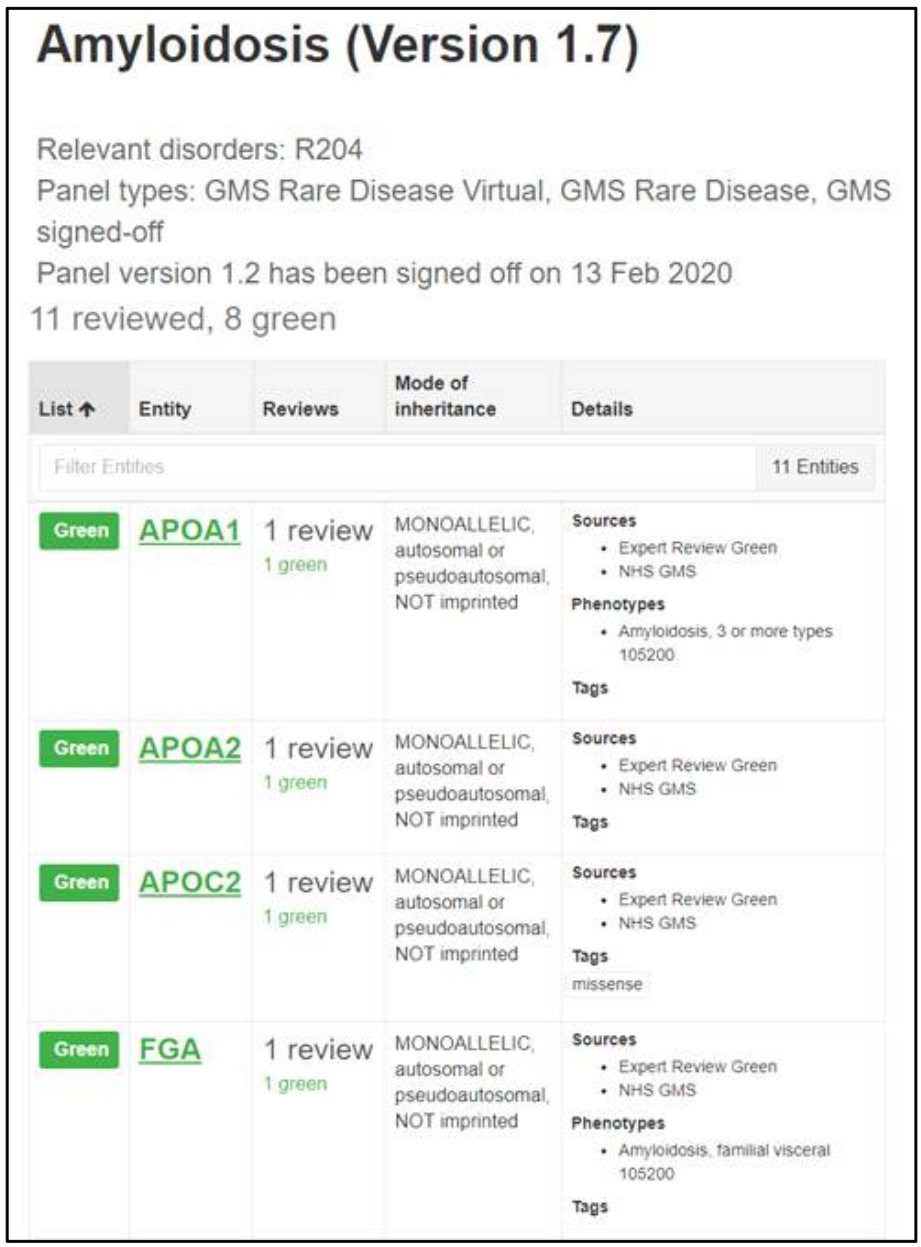
1) 전처리과정

가. 희귀질환-유전자

희귀질환과 관련된 유전자들의 정보들은 panel 웹사이트에서 희귀질환별로 다운로드를 받을 수 있다. 희귀질환별로 다운로드를 받을 경우, 빨간색, 노란색, 녹색 등급의 모든 유전자가 다 포함되어있다. panel에서는 희귀질환별 녹색 등급 유전자들만 별도로 다운로드를 받을 수 있게 되어있다. 따라서 희귀질환 315개에 대해 녹색 등급 유전자들만 다운로드 받은 뒤, 각각의 파일들을 병합해 희귀질환 이름과 유전자 목록을 테이블 형태로 만드는 과정이 필요하다. 희귀질환 이름별 유전자 목록을 만드는 이유는 유전자 관련 생물학적 패스웨이 검색을 위해 필요한 과정이다. 위와 같은 전처리과정을 통해 아래 표와 같은 해당 질병별 유전자 목록을 만든다. 각 질병별 파일에 들어가 보면, 유전자마다 설명과 함께 작성되어있다. 따라서 유전자 목록만 따로 가져오기 위한 작업을 파이썬으로 프로그래밍을 하여 작업을 하였다.

아래의 그림과 같이 Amyloidosis라는 질병을 찾아볼 경우, 녹색 등급으로 된 유전자에 대해 알 수 있고, 사이트의 아래쪽에 보면 녹색 등급만 다운로드를 받을 수 있게 되어있다. 전체로 다운로드를 받을 경우, 노란색 등급과 빨간색 등급에 대해서도 받을 수 있으나, 유전자들이 구분되기 힘들다. 질병 이름만 존재하고 아직 밝혀진 유전자가 없어 데이터가 없는 것도 있다. 그리고 Panel에서 제공하는 질병 이름은 HPO, OMIM 같은 질병코드 이름과는 다르기에 아직은 공통으로 쓰기 힘들다는 어려움이 있다.

아래 [그림6]은 PanelApp에서 Amyloidosis를 검색하였을 때 보여지는 유전자들의 형태와 전처리과정을 통해 변환되는 표 형태이다.



Amyloidosis (Version 1.7)

Relevant disorders: R204
Panel types: GMS Rare Disease Virtual, GMS Rare Disease, GMS signed-off
Panel version 1.2 has been signed off on 13 Feb 2020
11 reviewed, 8 green

List ↑	Entity	Reviews	Mode of inheritance	Details
Filter Entities				11 Entities
Green	APOA1	1 review 1 green	MONOALLELIC, autosomal or pseudoautosomal, NOT imprinted	**Sources** • Expert Review Green • NHS GMS **Phenotypes** • Amyloidosis, 3 or more types 105200 **Tags**
Green	APOA2	1 review 1 green	MONOALLELIC, autosomal or pseudoautosomal, NOT imprinted	**Sources** • Expert Review Green • NHS GMS **Tags**
Green	APOC2	1 review 1 green	MONOALLELIC, autosomal or pseudoautosomal, NOT imprinted	**Sources** • Expert Review Green • NHS GMS **Tags** missense
Green	FGA	1 review 1 green	MONOALLELIC, autosomal or pseudoautosomal, NOT imprinted	**Sources** • Expert Review Green • NHS GMS **Phenotypes** • Amyloidosis, familial visceral 105200 **Tags**

<패널(PanelApp) 화면상에 나오는 희귀질환별 유전자목록[3)]>

3) https://panelapp.genomicsengland.co.uk/panels/502/

	희귀질환 이름	유전자 목록
1	Additional findings reproductive carrier status	CFTR
2	Adult onset movement disorder	ACTB, AFG3L2, ANO3, APTX, ATM, ATP13A2, ATP1A2, ……
3	Adult solid tumours cancer susceptibility	APC, ATM, BAP1, BMPR1A, BRCA1, BRCA2, BRIP1, CBL, CDC73, ….
4	Adult solid tumours for rare disease	AIP, APC, ATM, BAP1, BMPR1A, BRCA1, BRCA2, BRIP1, CDC73,…
5	Albinism or congenital nystagmus	AP3B1, CACNA1A, CACNA1F, CASK, FRMD7, GPR143, HPS1….
6	Amelogenesis imperfecta	ACP4, AMBN, AMELX, C4orf26, CNNM4, COL17A1, DLX3,MMP20, …..
7	Amyloidosis	APOA1, APOA2, APOC2, FGA, GSN, LYZ, TTR, APOC3, B2M, CST3
8	Amyotrophic lateral sclerosis_motor neuron disease	ALS2, ANG, AR, DCTN1, FIG4, FUS, HNRNPA1, OPTN, …
9	Aniridia	FOXC1, ITPR1, PAX6, PITX2, ELP4, TRIM44, ISCA-37401-Loss
….	….	…
315	Xeroderma pigmentosum, Trichothiodystrophy or Cockayne syndrome	DDB2, ERCC1, ERCC2, ERCC3, ERCC4, ERCC5, ERCC6,…

[그림6] 아밀로도시스(Amyloidosis) 전처리과정 예시

나. 약물-유전자

약물에 관한 유전자 정보들은 Drugbank에서 다운로드를 하여 볼 수 있다. 해당 파일은 약물 1개당 유전자 1개로 되어있어, 약물 1개당 유전자 N개의 목록으로 변환할 필요가 있다. 파일의 총 길이는 10840개로 따라서 약물별로 유전자들을 그룹핑(Grouping) 하는 작업을 한다.[4] 유전자 목록을 만들어 관련 생물학적 패스웨이를 찾는데 필요한 작업이다. 유전자 목록을 만들기 위해서 약물과 유전자 목록 만드는 작업을 위해 파이썬 프로그래밍으로 판다스(pandas) 함수를 활용하였으며 유전자 그룹핑을 통해서 한 약물 당 얼마나 다양한 유전자가 연관되는지를 알 수 있다. 결과로 약물 2133개에 대해 유전자 2781개가 1개 이상으로 중복되어 관련되어있다는 것을 볼 수 있었다. 따라서 유전자 1개만 표적으로 약물이 나오는 경우는 많지 않다는 것을 알 수 있으며, 이에 따라서 유전자 간의 관계를 보기 위해 생물학적 패스웨이가 더 유용할 수 있다는 가능성을 볼 수 있다. Drugbank에 있는 약물들은 확실히 약물-단백질 관계가 밝혀진 것들만 존재하고 있다. 따라서 신약재창출을 위해서 이미 승인된 약물로 안정성이 확보되고 명확하게 약물과 유전자 간의 관계가 확실한 정보들이기에 본 논문 연구에 활용하는 것이 적합하다고 판단하였다. 이러한 이유로 Drugbank에 약물-유전자 관계를 전처리과정을 통해 얻은 정보를 활용하였다. [그림7]은 전처리과정을 통해 약물-유전자 관계가 어떻게 그룹핑 되는지를 보여준다.

4) Drugbank에서 제공하는 파일 다운로드
https://www.drugbank.ca/releases/latest#protein-identifiers

DrugBank ID	Drug Name	Type	UniProt ID	UniProt Name
DB00001	Lepirudin	BiotechDrug	P00734	Prothrombin
DB00002	Cetuximab	BiotechDrug	P00533	Epidermal growth factor receptor
DB00002	Cetuximab	BiotechDrug	O75015	Low affinity immunoglobulin gamma Fc region receptor III-B
DB00002	Cetuximab	BiotechDrug	P00736	Complement C1r subcomponent
DB00002	Cetuximab	BiotechDrug	P02745	Complement C1q subcomponent subunit A
DB00002	Cetuximab	BiotechDrug	P02746	Complement C1q subcomponent subunit B
DB00002	Cetuximab	BiotechDrug	P02747	Complement C1q subcomponent subunit C
DB00002	Cetuximab	BiotechDrug	P12318	Low affinity immunoglobulin gamma Fc region receptor II-a
DB00002	Cetuximab	BiotechDrug	P31994	Low affinity immunoglobulin gamma Fc region receptor II-b
DB00002	Cetuximab	BiotechDrug	P31995	Low affinity immunoglobulin gamma Fc region receptor II-c
DB00004	Denileukin diftitox	BiotechDrug	P01589	Interleukin-2 receptor subunit alpha
DB00004	Denileukin diftitox	BiotechDrug	P14784	Interleukin-2 receptor subunit beta
DB00004	Denileukin diftitox	BiotechDrug	P31785	Cytokine receptor common subunit gamma
DB00005	Etanercept	BiotechDrug	P01375	Tumor necrosis factor
...	...	...	...	...
DB15593	Golodirsen	BiotechDrug	P11532	Dystrophin

Drug Name	Uniprot ID
Lepirudin	P00734
Cetuximab	P00533,O75015,P00736,P02745,P02746,P02747,P08637, P09871,P12314,P12318,P31994,P31995
Denileukin diftitox	P01589,P14784,P31785
Abciximab	P05106,P08514,O75015,P00736,P02745,P02746,P02747, P08637,P09871,P12314,P12318,P31994,P31995,P04004
Amantadine	P21430,Q8TCU5,P14416,P36544,P43681,P32297
Benzatropine	P11229,Q01959,P35367,P31645,P23975
Cefazolin	P02918,P02919,P76577,P0AD65,P0AD68,P27169,P40933, P60568
Sorafenib	P35916,P04049,P35968,P15056,P36888,P09619,P10721,P 11362,P07949,P17948
Sildenafil	O76074,P18545,Q13956
Thalidomide	Q96SW2,P01375,P19838,P21802,P35354,P02763,P19652
Bosentan	P24530,P25101
Cinacalcet	P41180
Lomitapide	P55157
Pasireotide	P30872,P35346,P30874,P32745
...	...
Golodirsen	P11532

[그림7] 약물과 유전자의 관계를 위한 전처리과정

2) 희귀질환 - 약물의 공통 생물학적 패스웨이

희귀질환과 약물의 전처리과정으로 유전자 목록이 필요한 이유는, Reatome이란 생물학적 패스웨이 데이터베이스에서 제공하는 웹서비스인 분석 도구를 활용하기 위해서이다. 분석 도구는 유전자 목록을 넣으면 해당 생물학적 패스웨이를 FDR값을 기준으로 수치화해서 결과가 나온다. 따라서 Reatome 분석 도구 서비스를 활용하여 생물학적 패스웨이 정보를 얻기 위해서이다. Reatome에서는 분석방법을 RestfulAPI로 제공하여 프로그래밍을 통해 여러 질병이나 여러 약물에 대해 다양한 관련 유전자 목록을 검색할 경우, 해당 생물학적 패스웨이에 대한 정보를 수치화하여 찾을 수 있다. 따라서 희귀질환의 유전자 목록과 약물의 유전자 목록이 필요하다. 전처리과정을 끝낸 희귀질환과 약물의 유전자 목록을 분석서비스를 통해 결과물을 얻을 수 있다. 희귀질환-생물학적 패스웨이, 약물-생물학적 패스웨이 이렇게 각각 만들어진 목록에서 공통으로 사용된 생물학적 패스웨이만을 활용한다. 그 이유는 공통의 생물학적 패스웨이들만이 희귀질환과 약물 간의 관계를 더 확실하게 관련 있을 수 있기 때문이다. 희귀질환과 관련 있는 약물을 유전자를 공통의 관계로 보았을 때는 1570개의 약물이 나왔고, 희귀질환과 약물을 생물학적 패스웨이를 공통의 관계로 보았을 때 1883개가 나왔기 때문에 유전자로만 관계를 보는 것보다 생물학적 패스웨이를 활용하여 관계를 보는 것이 더 많은 정보를 포함하고 있다는 것을 알 수 있었다.

[그림8]은 희귀질환-생물학적패스웨이, 약물-생물학적패스웨이는 매트릭스(matirx)의 형태로 표현될 수 있으며, 생물학적 패스웨이를 기준으로 FDR값을 기준으로 벡터화를 통해서 유사도 계산이 가능하다. 기준이 되는 공통의 생물학적 패스웨이가 있어 희귀질환과 약물의 관계성을 설명하기 위한 활용이 가능하다.

[그림8] 희귀질환-패스웨이-약물의 관계 모식도

3) 거리 계산법

희귀질환-패스웨이, 약물-패스웨이 두 테이블의 유사도 확인을 위해서 FDR값을 벡터로 사용한다. FDR값을 사용하는 이유는 p-value값 보다 패스웨이가 좀 더 유의미하게 바라볼 수 있기 때문이다. [20]

FDR이란 다중비교문제에서 1종 오류를 조절하는 방법으로, 특히 유전학 연구에서 대량의 유전체 마커와 질병과의 연관성을 보는 연구에서 많이 사용되는 방법이다. 다중비교문제에서 기본적으로 많이 사용하는 본페로니 방법은 전체 테스트의 1종 오류를 alpha(예를들어 0.05)로 고정하는 방법이다. 즉, 전체 테스트가 유의하지 않은 경우, 유의하다고 잘못 판단할 확률을 0.05로 한 것이다. 수식은 아래와 같다. 따라서 유의하다고 판단한 것 중 틀릴 확률을 고정하는 새로운 p-value를 정의하는 방법으로 사용된다.

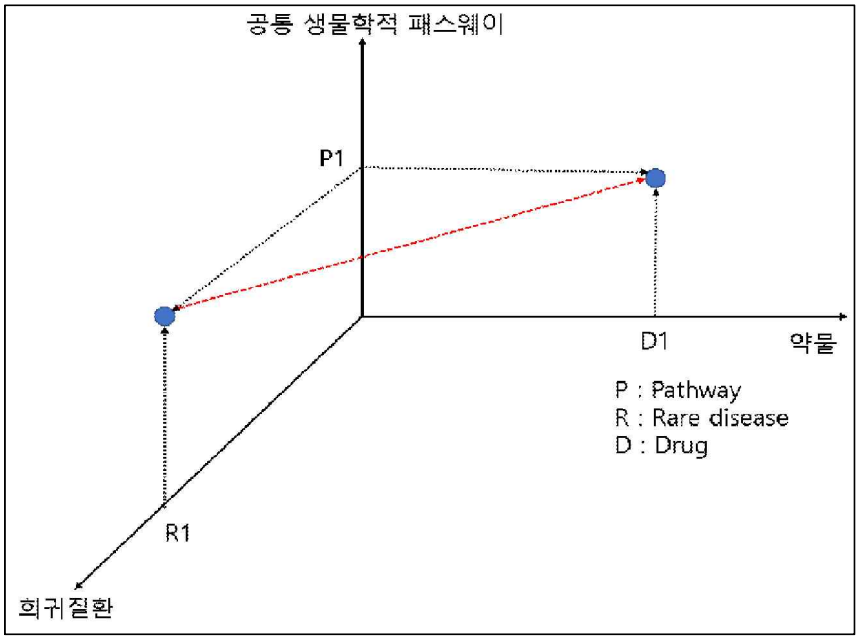
FDR = false positive / total positive
(total positive = false positive + true positive)

즉, 유의하다고 판단된 값 중 실제로는 유의하지 않은 것의 비율이다. 이 비율을 0.05로 고정한다면, 유의하다고 판단했을 때 틀릴 확률은 0.05로 고정할 수 있다. 실제 유의한(혹은 인과적 관계를 갖는) 마커를 알아야 FDR을 정확히 구할 수 있는데 이것은 어려운 일이다. 따라서 다양한 방법으로 FDR을 추정하는데 대표적인 방법으로 “Benjamini-Hochberg procedure”가 있다. 모든 마커의 p-value를 내림차순 정렬을 한 후, 개별 테스트의 p-value를 다르게 적용하는 방법으로 p-value가 낮은 마커로 갈수록 점점 엄격한 p-value의 컷오프를 적용하였다.

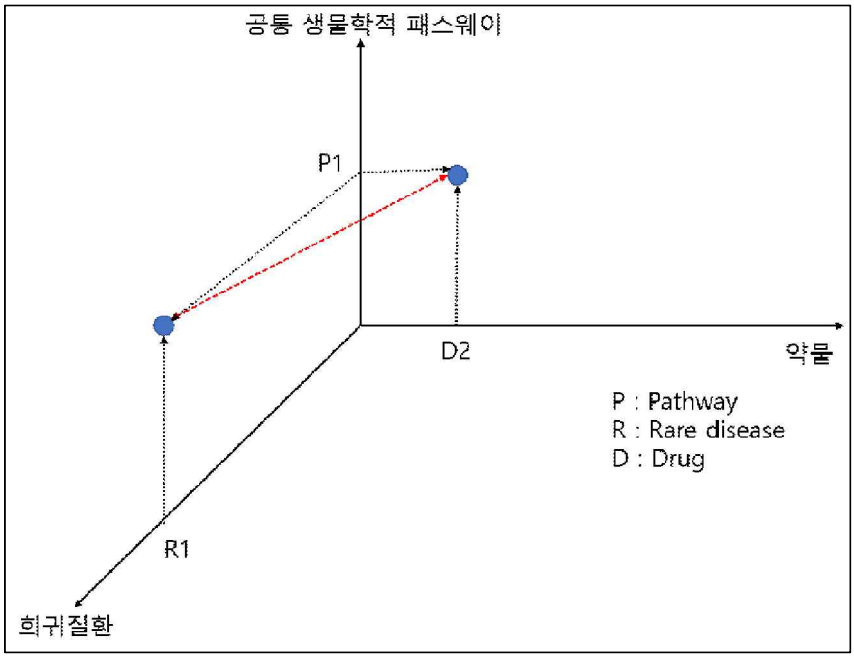
따라서 p-value는 유의미한 값을 알기 위해 쓰지만, 통계적으로 유의미한 결과는 찾았지만 실제로 그룹 평균에 차이가 없기에 다중 비교문제에서는 사용하기 어렵다. 따라서 여러 번의 테스트로 인해 문제를 극복

하는 방법 중에 한 종류인 FDR 값을 사용한다. 본 연구에서도 다중 비교문제인 여러 패스웨이들을 비교해서 보는 것이기 때문에, FDR 값을 활용하였다.

공통된 패스웨이가 기준이 되어 희귀질환과 약물이 계산될 수 있는 이유는 벡터화가 가능하기 때문이다. 벡터화를 하는 방식은 [그림9]와 [그림10]을 예시로 설명하였다. [그림9]은 D1과 R1에 대한 공통 패스웨이 P1으로 거리계산을 한 결과이고 [그림10]은 D2와 R1에 대한 공통 패스웨이 P1으로 거리계산 한 결과이다. 좌표값은 FDR값을 기준으로 한다. 공통된 패스웨이를 중심축이 되어, 희귀질환과 약물이 각각 한 축을 담당하게 된다. 희귀질환-패스웨이, 약물-패스웨이로 나오는 FDR 값을 기준으로 점을 찍으면 두 개의 점이 생긴다. 이 두 개의 점 사이의 거리를 구하면, 희귀질환과 약물의 거리를 측정할 수 있게 된다. D1과 D2의 비교를 통해서 거리계산을 통해 나온 값이 패스웨이 1개에 해당하는 거리가 측정된 것이다. 희귀질환과 약물의 거리 측정을 공통된 1883개의 패스웨이로 연산하기 위해 유클리디안 거리계산법이 사용된다.



[그림9] 예시. P1 기준 D1과 R1에 대한 관계 그래프



[그림10] 예시. P1 기준 D2와 R1에 대한 관계 그래프

공통된 패스웨이가 기준이 되기 때문에 유사도 계산을 할 수 있다. FDR 값들이 한 개의 점으로 존재하고, 각각의 파일에서 희귀질환 - 약물의 유사도는 유클리디안 유사도를 통해 계산한다. 유클리디안 유사도란, 유클리디안 거리를 구해서 두 벡터의 유사도로 사용하는 것을 말하며 계산법은 아래와 같다.

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

where $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n)$

희귀질환 - 약물의 유사도를 통해서 유사도 값이 작을수록 서로 가까운 거리에 존재한다는 것을 알 수 있다. 이러한 방법을 통해서 희귀질환의 신약재창출 후보 가능성을 수치화하여 볼 수 있고, 희귀질환과 관련된 약물을 찾을 수 있다.

Ⅲ. 결 론

1. 희귀질환-약물 유사도

희귀질환과 약물의 유사도를 가진 데이터 테이블을 만들었다. 희귀질환-약물 간의 관계를 FDR 값을 유사도 계산을 통해 만들어진 결과가 아래와 같은 데이터 테이블의 형태로 나온다. 질병 309개에 대한 약물 1888개에 대한 유사도 결과이다. 희귀질환별로 약물의 리스트, 약물별 희귀질환 리스트를 만들어서 확인할 수 있다.

Drug / Disease_name	Bortezomib	Sildenafil	Sorafenib	Bosentan	Azacitidine	...	Vigabatrin
Mitochondrial disorder with complex II deficiency	0.10280217	0.5409945	1.897736076	0.014810164	0.531099352	...	0.372542093
Erythropoietic protoporphyria, mild variant	0.103071139	0.541045675	1.897750665	0.016574518	0.531151481	...	0.372616405
Mitochondrial disorder with complex III deficiency	0.107007993	0.541809444	1.897968556	0.033192853	0.531929456	...	0.373724545
Mitochondrial disorder with complex V deficiency	0.1092089	0.542248418	1.898093915	0.039719502	0.532376577	...	0.374360667
Leber hereditary optic neuropathy	0.19104211	0.565218121	1.904783108	0.164364888	0.55575444	...	0.406920246
Cardiac arrhythmias - additional genes	0.634243291	0.83593624	1.936346064	0.637443087	0.778871291	...	0.738174748
Combined factor V and VIII deficiency	0.65374069	0.850420579	2.007969703	0.656322675	0.766432475	...	0.754538087
Severe familial anorexia	0.950978851	1.115641548	1.923351846	0.97580754	1.07335552	...	1.044398691
Bladder cancer pertinent cancer susceptibility	1.355875203	1.472566107	2.010852816	1.369669753	1.415391587	...	1.419353199
Upper gastrointestinal cancer pertinent cancer susceptibility	1.355875203	1.472566107	2.010852816	1.369669753	1.415391587	...	1.419353199
Hydroa vacciniforme	1.452623662	1.56433161	2.39913746	1.467882038	1.404842906	...	1.514346707
Familial disseminated superficial actinic porokeratosis	1.523753683	1.653670372	2.458323181	1.562744378	1.583197815	...	1.606467635
Additional findings reproductive carrier status	1.544297254	1.662324527	2.35671683	1.57189921	1.615686416	...	1.615374692
...	...	...	...	...	...	...	...
Idiopathic ventricular fibrillation	1.634616263	1.722421906	2.169495824	1.635323061	1.71933961	...	1.658945503

[표4] 희귀질환-약물 유사도 결과값

2. 신약재창출 약물 후보

가. FDA승인된 희귀질환 치료제와의 비교

위의 데이터 테이블의 정보가 타당한지 알기 위해 FDA 승인을 받은 희귀질환 치료제를 검색하여 출력되는지를 확인한다. FDA 승인받은 희귀질환 약에 대한 정보는 FDA 공식 홈페이지[5]에서 구할 수 있으며, 해당 목록을 다운로드를 받은 뒤, 해당 약물과 Drugbank의 승인된 약물과의 공통 약물만을 추출한다.

이들 중 FDA 승인이 되었으면서, 희귀질환에 후보가 될 수 있는 약들에 대해 어떻게 결과가 나오는지에 대해 살펴보려고 한다. 유사도 값의 평균은 11로 최대 42 이하의 값으로 유사도 값은 도출된다. 값의 형태이기 때문에 순위를 매길 수 있다.

약물을 기준으로 할 경우, 희귀질병 309개에 대해 순위를 매기를 매길 수 있다. 어떤 특정 희귀질환이 어떤 약이 더 잘 들지에 대해서는 추측이 많아, 이미 치료제로 쓰이고 있는 약과 그 치료제와 유사한 질병을 찾아보는 방식으로 앞으로 계산한 결과가 얼마나 의미가 있는지에 대하여 확인하였다.

승인된 약 중 4가지를 채택한 것은 표적 전략으로 만들어진 희귀질환 관련 약으로 승인된 약을 기준으로 하였다.[21] Bosetan, Bortezomib, Sorafenib, Lomitapide 4가지 약 중에서 Lomitapide가 가장 희귀질환과 가까운 곳에 있었다. 따라서 어떻게 가깝게 나올 수 있었는지 살펴보았다.

---

5) FDA 승인받은 희귀질환 치료제
https://www.accessdata.fda.gov/scripts/opdlisting/oopd/

[FDA승인되어 희귀질환 치료제로 쓰이는 약 정보와의 결과 값 비교]

DrugName	Year	Discovery strategy	Therapeutic area	Indication	Molecular mechanism of action	Similarity value	Rank	Panel_rare_disease_name
Bosentan	2001	Target	Cardiovascular	Pulmonary arterial hypertension	Endothenlin-1 receptor antagonists	**5.079826**	**84**	**Pulmonary arterial hypertension**
Bortezomib	2003	Target	Cancer	Multiple myelomas	Proteasome	6.285197	113	Inherited predisposition to acute myeloid leukaemia (AML)
Sorafenib	2005	Target	Cancer	Renal cell carcinoma	Multikinase inhibitor VEGFR/PDGFR/KIT	6.008331	110	**Renal cancer pertinent cancer** susceptibility
Lomitapide	2012	Target	Metabolic	Homozygous familial hypercholesterolemia	Binds to microsomal triglyceride transfer protein	**2.816444**	**34**	**Familial hypercholesterolaemia - targeted panel**

Value mean = 11

Ranking max= 309(희귀질환 총 개수)

[표5] FDA 승인되어 희귀질환 치료제로 쓰이는 약 정보와의 결과 비교

나. Thalidomide

Thalidomide는 1950년대 후반~1960년대까지 임산부들의 입덧 방지용으로 판매된 약으로 부작용으로 기형아들이 출산 되어 사용이 금지되었다. 하지만 최근, 신생혈관 억제라는 탈리도마이드 부작용이 한센병, 다발성 골수종, 암 등의 치료에 쓰일 수 있다는 것이 알려져 제한된 경우 사용을 할 수 있게 되었다.

희귀질환과 약물을 공통 생물학적 패스웨이로 연결하였을 때와 희귀질환과 약물을 공통 유전자로 연결하였을 때의 차이를 알아보고 그 차이를 통해서 희귀질환과 약물의 관계를 패스웨이로 연결한 정보가 정말로 신뢰할 수 있는지에 대해 확인한다.

본 연구 방법을 통해 “패스웨이” 중심으로 희귀질환-패스웨이-약물로 목록을 만들었을 때, 약물 1888개에 대한 정보가 있다. 하지만 “유전자” 중심으로 희귀질환-유전자-약물로 공통 유전자만으로 찾았을 때 약물 1570개의 정보가 나온다. 따라서 희귀질환-패스웨이-약물의 데이터가 더 풍부한 자료를 가지고 있음을 알 수 있다. 또한, 얼마나 유사한지에 대해 순위를 통해 알 수 있어 유용하다. 해당 내용은 희귀질환-패스웨이-약물과 희귀질환-유전자-약물을 Thalidomide 약물로 비교한 내용을 설명한다.

희귀질환-패스웨이-약물 후보목록에서 Thalidomide를 검색하였을 경우 나오는 리스트이다. 질병 309개에 대해 유사도 값으로 순서가 매겨져 얼마나 유사한지에 대한 정보를 알 수 있다.

또한 Thalidomide 약이 Bladder cancer의 치료제가 될 수 있다는 논문을 찾아볼 수 있었다.[22] 따라서 약물-희귀질환의 관계를 생물학적 패스웨이로 봐도 가능성이 있다는 것을 알 수 있다.

희귀 질환 이름 Panel	유사도 결과	순위
Bladder cancer pertinent cancer susceptibility	2.939787253	1
Upper gastrointestinal cancer pertinent cancer susceptibility	2.939787253	2
Prostate cancer pertinent cancer susceptibility	2.959662566	3
Severe familial anorexia	2.965537474	4
Classical tuberous sclerosis	3.011405328	5
Familial hidradenitis suppurativa	4.298690827	47
Familial Meniere Disease	19.72354457	261
Arthrogryposis	25.37937279	284
Autism	33.70377429	302
DDG2P	39.44676888	305

[표6] Thalidomide와 관련 있는 희귀질환 순위

희귀질환 이름	유전자	약물 이름	Uniprot_ID
Arthrogryposis	FGFR2	Thalidomide	P21802
Autism	PTGS2	Thalidomide	P35354
DDG2P	CRBN	Thalidomide	Q96SW2
Familial hidradenitis suppurativa	TNF	Thalidomide	P01375
Familial Meniere Disease	NFKB1	Thalidomide	P19838

[표7] 희귀질환-유전자-Thalidomide(약물)로 봤을 때 결과

다. Lomitapide

Lomitapide는 미국에서 상품명 Juxtapid 로, EU에서 상품명 Lojuxta 로 판매되는 Aegerion Pharmaceuticals 회사에서 개발한 가족성 고콜레스테롤 혈증 치료를 위한 지질 강하제로 사용되는 약이며, 2012년에 희귀질환 치료제로 승인받았다.

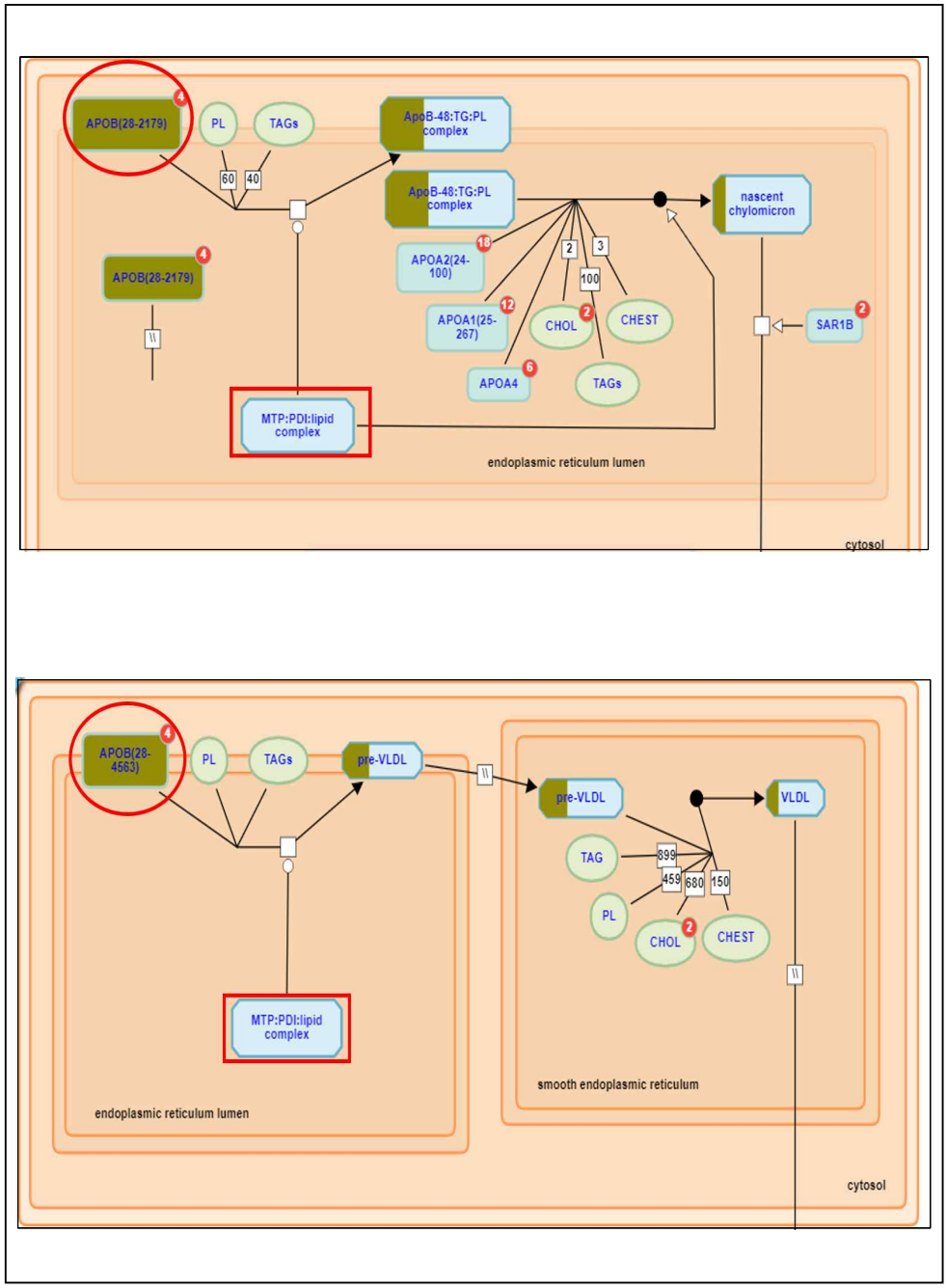
panel 데이터베이스에서 가족성 고콜레스테롤 희귀질환의 이름은 “Familial hypercholesterolaemia - targeted panel”로 되어있다. 희귀질환의 치료제로 사용되고 있는 Lomitapide와 희귀질환 “Familial hypercholesterolaemia - targeted panel” 의 관계를 통해서 본 연구의 희귀질환-약물 후보 순위가 정말로 의미가 있는지를 확인해보았다. Lomitapide를 기준으로 희귀질환 309개를 정렬하였을 때 “Familial hypercholesterolaemia - targeted panel” 은 33번째의 순위로 나왔으며, 유사도 값은 2.816443704로 나왔다. 따라서 관련이 높게 나왔음을 알 수 있었다. 관련이 높게 나올 수 있었던 이유를 약물-유전자, 희귀질환-유전자를 어떻게 패스웨이로 연결되는지를 Reatome과 Bee 프로그램을 통해서 살펴보았다.

Lomitapide는 Drugbank 기준으로 표적 유전자가 “MTTP” 이고, “Familial hypercholesterolaemia - targeted panel”은 panel 기준으로 희귀질환 관련된 유전자는 “APOB, APOE, LDLR, LDLRAP1, PCSK9” 이다. 이 둘 사이의 관계가 유전자로만 바라보았을 때는 전혀 관련 없어 보이지만, 생물학적 패스웨이 경우에는 관련성이 높다는 것을 알 수 있다.

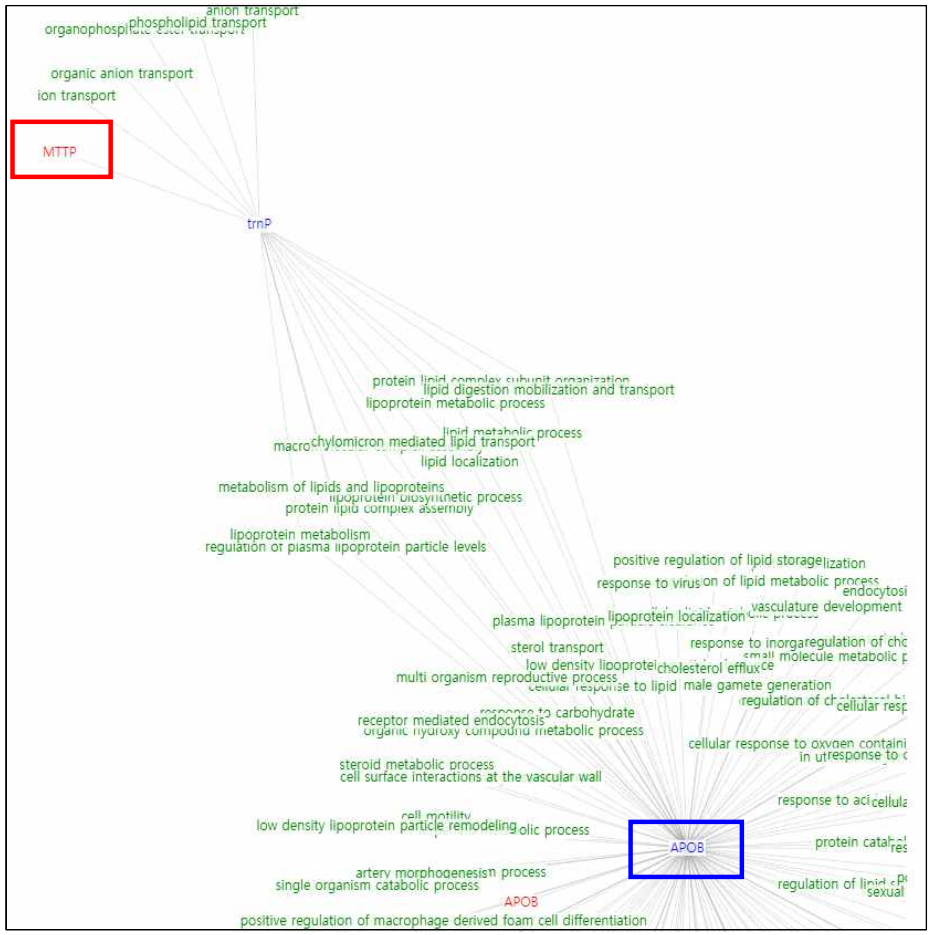
[그림11]은 Reatome에서 유전자 “MTTP”와 “APOB” 의 관계를 보여준다. 빨간색 동그라미 표시가 “APOB”이고 빨간색 네모 표시가 “MTTP”를 말한다. [그림12]은 Bee 프로그램을 통해서 “MTTP”와 “APOB”의 관계를 패스웨이로 어떻게 연결되는지 보여준다.

	희귀질환 이름	Lomitapide
1	Mitochondrial disorder with complex II deficiency	0.065528407
2	Erythropoietic protoporphyria, mild variant	0.06594957
3	Mitochondrial disorder with complex III deficiency	0.07194718
4	Mitochondrial disorder with complex V deficiency	0.07518158
5	Leber hereditary optic neuropathy	0.176324836
6	Cardiac arrhythmias - additional genes	0.640631189
7	Combined factor V and VIII deficiency	0.659419506
8	Severe familial anorexia	0.977893137
...	...	...
33	**Familial hypercholesterolaemia - targeted panel**	**2.816443704**
...	...	...

[표8] Lomitapide에 대한 희귀질환별 유사도 값



[그림11] “Reactome”에서 검색했을 때, MTTP와 APOB의 관계



[그림12] BEE에서 관련 유전자 간의 관계를 보여주는 네트워크

라. Bosentan

Bosentan은 상품명 “Tracleer”로, 폐동맥 고혈압의 치료에 사용되는 이중엔도 텔린 수용체 길항제이다. “Pulmonary arterial hypertension”이라는 희귀질환의 치료제로 사용된다. Botentan은 “EDNRB, EDNRA”을 표적 유전자로 만들어졌다. 희귀질환 “Pulmonary arterial hypertension”은 panel에서 관련 유전자로 “ACVRL1, ATP13A3, BMPR2, EIF2AK4, ENG, GDF2, KCNK3, SMAD9, SOX17, TBX4, AQP1, CAV1, BMPR1B, CBLN2, KCNA5, SARS2, SMAD1, SMAD4”가 관련 있음을 알 수 있다. 유전자로만 봤을 때, 둘 사이의 관계는 겹치는 부분이 없다. 하지만 생물학적 패스웨이로 보았을 때는 둘 사이의 관계가 존재하였고, 아래의 표에서처럼 309개의 약 중에 84번째로 “5.046478409”라는 유사도 값으로 유의미한 결과를 볼 수 있다.

순위	희귀질환 이름	Bosentan
1	Mitochondrial disorder with complex II deficiency	0.10280217
2	Erythropoietic protoporphyria, mild variant	0.103071139
3	Mitochondrial disorder with complex III deficiency	0.107007993
4	Mitochondrial disorder with complex V deficiency	0.1092089
5	Leber hereditary optic neuropathy	0.19104211
6	Cardiac arrhythmias - additional genes	0.634243291
	...	...
83	Hypophosphataemia or rickets	5.045903899
84	**Pulmonary arterial hypertension**	**5.046478409**
...	...	...

[표9] Bosentan 약물에 대한 희귀질환별 유사도 값

마. Bortezomib

상품명 Velcade로 판매되는 Bortezomib은 다발성 골수종 및 외투 세포 림프종을 치료하는 데 사용되는 항암제이다.

순위	희귀질환 이름	Bortezomib
1	Mitochondrial disorder with complex Ⅱ deficiency	0.10280217
2	Erythropoieticprotoporphyria, mild variant	0.103071139
3	Mitochondrial disorder with complex Ⅲ deficiency	0.107007993
4	Mitochondrial disorder with complex V deficiency	0.1092089
5	Leberhereditary optic neuropathy	0.19104211
6	Cardiac arrhythmias - additional genes	0.634243291
7	Combined factor V and Ⅷ deficiency	0.65374069
8	Severe familial anorexia	0.950978851
...	...	...
113	**Inherited predisposition to acute myeloid leukaemia (AML)**	**6.249022673**
...	...	...

[표10] Bortezomib 약물에 대한 희귀질환별 유사도 값

Bortezomib은 “PSMB5, PSMB1”를 표적 유전자로 갖고 있으며, 다발성 골수종 치료제로 panel에서 희귀질환 이름은 “Inherited predisposition to acute myeloid leukaemia (AML)”으로 “ANKRD26, CEBPA, DDX41, ETV6, GATA2, RUNX1, TERC, TERT, TP53, ACD, CHEK2, RTEL1, SAMD9, SRP72” 의 유전자들을 포함하고 있다. 유전자로만 바라보았을 때, 둘 사이의 관계성을 찾기 어려웠지만, 생물학적 패스웨이를 통해서 위의 표처럼 질병 309개 중에서 113번째에 “6.249022673” 유사도 값으로 유의미한 관계를 볼 수 있었다.

Ⅳ. 고 찰

본 논문의 연구 방법론은 희귀질환에 대한 신약재창출 후보 목록을 볼 수 있고 얼마나 연관성이 있는지에 대해 그 후보목록을 순위별로 볼 수 있다는 장점이 있다. 따라서 다른 질병도 유전자정보와 생물학적 패스웨이를 활용하여 신약재창출 후보를 순위별로 알 수 있다는 것을 기대할 수 있다. 신약재창출 실험연구를 진행하기 전에 그 약물에 대한 질병 후보 목록, 또는 질병에 대한 약물 후보목록을 알아보고 신약재창출의 목표선정을 도와줄 수 있다.

본 논문에서는 승인된 약물에 대한 유전자정보만을 사용하였는데, 약물-유전자 간의 관계를 나타낸 데이터베이스는 여러 개 존재한다. 그중에서 DGI(Drug-Gene-Interaction) 데이터베이스에는 승인된 약물에 대해 더 많은 유전자에 대한 정보가 존재한다. 하지만 DGI 데이터베이스에서 승인된 약물의 개수는 1445개로 DrugBank는 1888개로 차이가 있었다. 본 연구의 확장으로, 패스웨이 분석을 하였을 때는 DGI에서는 1943개의 패스웨이가 존재해 Drugbank와 DGI 데이터베이스를 비교분석을 하거나, 통합하여 희귀질환 패스웨이와 비교해 볼 수 있다.

생물학적 패스웨이 데이터베이스도 다양하게 존재하고 있다. 따라서, 본 연구에서 사용한 Reatome이 아닌 생물학적 패스웨이 데이터베이스를 사용하였을 때, 어떻게 결과가 나오는지에 대한 확인도 필요하다. 생물학적 패스웨이는 각각 데이터베이스마다 데이터의 양의 차이도 크기 때문에 연구의 데이터 확장을 위해 필요한 정보이다. 또한, “Wikipathways” 생물학적 데이터베이스에서는 희귀질환 관련된 패스웨이가 따로 존재한다. 희귀질환 패스웨이를 활용하였을 때의 결과는 어떻게 나올 수 있는지에 대한 연구가 되어 현재 유전자를 통해 분석한 값으로 나온 생물학적 패스웨이들과 비교분석이 필요하다.

본 논문에서 쓰인 희귀질환-패스웨이-약물 후보목록에서 어떤 패스웨

이에서 어떤 유전자로 인해 이렇게 관련성이 있는지에 대해 볼 필요가 있다. 관련 유전자에 가중치를 두고 유사도 계산을 한다면 보다 정확한 정보를 얻을 수 있지 않을까 가설을 세워봤었다. 따라서 약물로 유의해서 볼 때는 약물 관련 유전자로, 희귀질환 유전자에 해당 유전자에 대해 가중치를 방법을 활용하여 희귀질환-약물의 생물학적 패스웨이 유사도 신뢰도를 더 높일 수 있다고 생각하였다.

희귀질환에 대한 치료제 연구에 더 많은 발전을 위해서는 사람들이 좀 더 관심을 기울이고 적극적으로 관련 연구를 위해 힘쓸 필요도 있다. 데이터가 한정적인 만큼, 최소한의 데이터로 최대한의 정보를 얻을 수 있도록 다양한 시도가 필요하다고 생각되었다. 이번 연구가 앞의 다양한 시도 중에 한 방법으로 제안되었다고 생각한다. 마무리하면서, 앞으로 더 많은 희귀질환 치료제가 나오길 바란다.

참 고 문 헌

〔1〕 희귀질환 범위 늘리고, 의료비 등 지원확대 관련기사
https://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04&MENU_ID=0403&page=1&CONT_SEQ=346076

〔2〕 Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs, Nat Rev Drug Discov, 2004, vol. 3 (pg. 673-83)

〔3〕 Metz JT, Hajduk PJ. Rational approaches to targeted polypharmacology: creating and navigating protein-ligand interaction networks, Curr Opin Chem

〔4〕 Boran AD, Iyengar R. Systems approaches to polypharmacology and drug discovery, Curr Opin Drug Discov Devel, 2010, vol. 13 (pg. 297-309)

〔5〕 Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., … Pirmohamed, M. (2018). Drug repurposing: Progress, challenges and recommendations. Nature Reviews Drug Discovery, 18(1), 41 - 58. https://doi.org/10.1038/nrd.2018.168

〔6〕 Xue, H., Li, J., Xie, H., & Wang, Y. (2018). Review of drug repositioning approaches and resources. International Journal of Biological Sciences, 14(10), 1232 - 1244. https://doi.org/10.7150/ijbs.24612

[7] Lotfi Shahreza, M., Ghadiri, N., Mousavi, S. R., Varshosaz, J., & Green, J. R. (2018). A review of network-based approaches to drug repositioning. Briefings in Bioinformatics, 19(5), 878 – 892. https://doi.org/10.1093/bib/bbx017

[8] Sun, Y., Sheng, Z., Ma, C., Tang, K., Zhu, R., Wu, Z., … Cao, Z. (2015). Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer. Nature Communications, 6, 1 – 10. https://doi.org/10.1038/ncomms9481

[9] Oprea TI, Overington JP. Computational and Practical Aspects of Drug Repositioning. Assay Drug Dev Technol. 2015;13:299 – 306.

[10] Sun, Y., Sheng, Z., Ma, C., Tang, K., Zhu, R., Wu, Z., … Cao, Z. (2015). Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer. Nature Communications, 6, 1 – 10. https://doi.org/10.1038/ncomms9481

[11] Palma G, Vidal M-E, Raschid L. Drug-target interaction prediction using semantic similarity and edge partitioning. ISWC. 2014;1:131 – 46.

[12] Park, K. (2019). A review of computational drug repurposing. Translational and Clinical Pharmacology, 27(2), 59 – 63. https://doi.org/10.12793/tcp.2019.27.2.59

[13] Emig, D., Ivliev, A., Pustovalova, O., Lancashire, L., Bureeva, S., Nikolsky, Y., & Bessarabova, M. (2013). Drug Target Prediction and

Repositioning Using an Integrated Network-Based Approach. PLoS ONE, 8(4). https://doi.org/10.1371/journal.pone.0060618

[14] Jadamba, E., & Shin, M. (2016). A Systematic Framework for Drug Repositioning from Integrated Omics and Drug Phenotype Profiles Using Pathway-Drug Network. BioMed Research International, 2016. https://doi.org/10.1155/2016/7147039

[15] Mejía-Pedroza, R. A., Espinal-Enríquez, J., & Hernández-Lemus, E. (2018). Pathway-based drug repositioning for breast cancer molecular subtypes. Frontiers in Pharmacology, 9(AUG), 1 - 13. https://doi.org/10.3389/fphar.2018.00905

[16] Iwata, M., Hirose, L., Kohara, H., Liao, J., Sawada, R., Akiyoshi, S., … Yamanishi, Y. (2018). Pathway-Based Drug Repositioning for Cancers: Computational Prediction and Experimental Validation. Journal of Medicinal Chemistry, 61(21), 9583 - 9595. https://doi.org/10.1021/acs.jmedchem.8b01044

[17] Martin, A. R., Williams, E., Foulger, R. E., Leigh, S., Daugherty, L. C., Niblock, O., … McDonagh, E. M. (2019). PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nature Genetics, 51(11), 1560 - 1565. https://doi.org/10.1038/s41588-019-0528-2

[18] Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., … Wilson, M. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Research, 46(D1),

D1074 - D1082. https://doi.org/10.1093/nar/gkx1037

〔19〕 Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., … D'Eustachio, P. (2018). The Reactome Pathway Knowledgebase. Nucleic Acids Research, 46(D1), D649 - D655. https://doi.org/10.1093/nar/gkx1132

〔20〕 FDR 값에 대한 설명, 참고 https://3months.tistory.com/262

〔21〕 Swinney, D. C., & Xia, S. (2014). The discovery of medicines for rare diseases. Future Medicinal Chemistry, 6(9), 987 - 1002. https://doi.org/10.4155/fmc.14.65

〔22〕 Huang, Y. T., Cheng, C. C., Chiu, T. H., & Lai, P. C. (2015). Therapeutic potential of thalidomide for gemcitabine-resistant bladder cancer. International Journal of Oncology, 47(5), 1711 - 1724. https://doi.org/10.3892/ijo.2015.3155

Abstract

A Study on Drug Repositioning for Rare Diseases based on Biological Pathways

Hyeyoen Kim
Healthcare Management and Informatics
The Graduate School
Seoul National University

Introduction: The purpose of this study is to utilize biological pathway data for rare disease drug repositioning. There are more than 7,000 rare diseases worldwide, but there is only treatment for 5% of these diseases. While there is a great need for treatments, traditional drug development is a very time consuming and costly process. For rare disease treatment, drug repositioning can potentially be a quicker and cheaper alternative. Biological pathway data describe the interaction between biological elements in detail and can be used to analyze gene data from a wider perspective. Therefore, it is hypothesized that they are suitable to use in drug repositioning. In this study, a common biological pathway list is generated from drug-related and rare disease-related gene data to find new drug candidates for rare diseases. Using the common pathway list and rare disease-drug similarity, a rare disease-drug candidate list is generated.

Methods: 309 rare diseases from the Genomics England PanelApp is utilized with the relevant genes. 1,888 approved drugs and related genetic information is used from DrugBank. Using analysis tools provided by Reactome, biological pathways relevant to the rare disease-gene and drug-gene lists were collected based on FDR values. Among the collected biological pathways, there are 1,883 biological pathways commonly associated with the rare diseases and drugs, which are then used to calculate the similarity between the rare diseases and drugs. The Euclidean similarity of the rare diseases and drugs are calculated by vectorizing the FDR values.

Results: Through this study, a rare disease-drug candidate list was generated. In the list, it can be interpreted that the smaller the value between a rare disease and drug is the more similar they are. Therefore, the more similar a rare disease and drug is, the more likely it is to be a candidate for rare disease drug repositioning. The results were compared with existing approved drugs used to treat rare diseases, for evaluation. Lomitapide is a drug used to treat "Homozygous familial hypercholesterolemia". In the drug-rare disease list it has a similarity value of 2.8 with its PanelApp equivalent disease, which is rank 34 out of 309 rare diseases. The rare disease-pathway-drug results were also compared with the rare disease-gene-drug results with the drug, Thalidomide. In the rare disease-pathway-gene results, it is observed that "Bladder cancer pertinent cancer susceptibility" was the closest disease to Thalidomide, which coincides with recent literature.

Discussion: From the results, it can be confirmed that the rare disease-drug list was relevant with existing rare disease treatments

and that this relevance can also be measured. In addition, it is found that rare disease-pathway-drug associations are more applicable to drug repositioning than rare disease-gene-drug associations. Finally, it is believed that biological pathways can be used not just for rare diseases but also for finding drug repositioning candidates in common diseases.

keywords : Rare disease, Orphan disease, drug discovery, drug repositioining, biological pathway

Student Number : 2017-27136