Journal of Biomedical Engineering Research 43: 230-240 (2022)
http://dx.doi.org/10.9718/JBER.2022.43.4.230

학 술 논 문

분자구조 유사도를 활용한 약물 효능 예측 알고리즘 연구

정화영* · 송창현* · 조혜연 · 기재홍‡

연세대학교 의공학부

A Study on the Prediction of Drug Efficacy by Using Molecular Structure

Hwayoung Jeong*, Changhyeon Song*, Hyeyoun Cho and Jaehong Key‡

Department of Biomedical Engineering, Yonsei University
(Manuscript received 19 July 2022 ; revised 9 August 2022 ; accepted 10 August 2022)

Abstract: Drug regeneration technology is an efficient strategy than the existing new drug development process, which requires large costs and time by using drugs that have already been proven safe. In this study, we recognize the importance of the new drug regeneration aspect of new drug development and research in predicting functional similarities through the basic molecular structure that forms drugs. We test four string-based algorithms by using SMILES data and searching for their similarities. And by using the ATC codes, pair them with functional similarities, which we compare and validate to select the optimal model. We confirmed that the higher the molecular structure similarity, the higher the ATC code matching rate. We suggest the possibility of additional potency of random drugs, which can be predicted through data that give information on drugs with high molecular similarities. This model has the advantage of being a great combination with additional data, so we look forward to using this model in future research.

Key words: SMILES, LINGO similarity, ATC code, Drug similarity, Drug regeneration

I. 서 론

신약개발은 인류의 건강한 삶을 위해 지속적인 노력이 요구되는 연구분야이다. 하지만 신약개발은 많은 비용과 개발기간, 연구 인력이 필요하다. 이러한 비용의 감소를 위해 최근 빅데이터와 딥러닝 및 머신러닝을 이용한 In silico기반 신약설계 기술이 다양하게 연구되어 제시되고 있다. 이러한 배경에서 화합물 분자구조를 문자열로 표기할 수 있는 Simplified molecular-input line-entry system (SMILES) 방법이 주목을 받는다[1]. SMILES는 ASCII 문장으로 화합물의 구조를 나타내기 위한 표기법이다. SMILES는 최소의 간단한 규칙만을 이용해 2D 분자구조를 문자열로 변환하는 과정에서 분자구조적 특징을 포함할 수 있는 압축적인 방식이므로 정보처리 측면에서도 매우 효율적이다. 이와 같은 장점으로 최근 CNN (Convolutional Neural Network)을 이용하여 제시된 2D 분자구조 이미지를 SMILES 문자열로 자동변환해주는 알고리즘의 개발이 활발하게 진행되고 있다[2,3].

기존의 전통적 신약개발연구는 질병의 작용점을 선정한 후, 약물 스크리닝을 통한 약물의 최적화 등의 단계로 이루어진다. 전임상 과정, 임상 과정을 1, 2, 3단계로 나누어 진행하며 도출된 결과는 최종적으로 FDA 심사와 등록 개발단계를 거친다. 위 과정은 평균적으로 10-15년의 시간과 10억 달러 이상의 막대한 자금이 소요된다. 그러나 이러한 투자에도 불구하고 신약개발의 가능성은 10% 내외이다[4]. 신약재창출은 이미 시장에서 판매를 하고 있거나 산업화에 실패한 약물의 새로운 의학적인 용도를 개발하는 신약개발의 방법이다. 이는 안정성이 이미 검증된 약물을 활용하여 비용을 절감할 수 있으며, 개발 기간을 단축시킬 수 있다는 장점을 가진다.

따라서 본 연구에서는 신약개발의 한 측면으로 신약재창출의 중요성을 인지하고 그것을 중점으로 시장에서 판매되는 약물들 사이에 약리학적 기능의 유사성을 검출하여 기존 약물의 신

*Corresponding Author : Key, Jaehong
Department of Biomedical Engineering, Yonsei University, 1 Yonseidae-gil, Wonju, Gangwon-do, 220-710, South Korea
Tel: +82-33-760-2587
E-mail: jkey@yonsei.ac.kr
‡Contributed equally to this work

기능 발견 가능성을 제시하는 것에 목표를 둔다. 각 약물은 그들의 특성에 따라 흡수, 이동, 합성, 분해, 대사, 배출 등의 생리 작용이 다르게 이루어지므로 이들을 상징하고 표시하는 방법이 필요하다. 분자구조는 이들이 공통적으로 갖고 있는 가장 기본적인 단위이므로 약물 분류 목적에 적합할 뿐 아니라 활용 시 가치가 높다. 본 연구에서는 이 점에 착안하여 주 데이터를 약물의 분자구조로 선택하고, 이들과 약리학적 기능을 쌍으로 연결한다. 약물 간 분자 구조의 유사도를 약리학적 기능의 유사성과 관련 지을 수 있다면, 기존 약물의 재평가 및 식용으로 검증된 천연물, 화합물에 대해서도 약리학 관점의 새로운 가이드라인을 제시할 수 있다.

본 연구에서는 '분자구조가 유사하다면 기능적 유사도가 높을 것이다'를 가설로 설정한다. 11,281개의 약물을 SMILES 표기법으로 분류하고 4가지(Edit distance(Ed), NLCS distance, LINGO similarity, LINGO cosine similarity) 문자열 검색 알고리즘을 이용하여 구조적 유사도를 산출한다. 구현된 유사도 모델의 정확성을 검증하고 최적화 모델을 선별하기 위해 ATC code를 기능적 분류 기준(Gold Standard)으로 사용한다[5]. 나아가 제시한 모델을 통해 분자구조의 유사성 정보가 체내 기능의 유사성을 예측하는 데 도움을 줄 수 있다는 것을 검증하고, 향후 신약재창출의 기반이 될 수 있는 가능성을 기존 약물과 천연물을 통해 제시한다.

II. 연구방법

1. SMILES

SMILES는 화학 물질의 구조를 ASCII 문장으로 나타내는 방법 중 하나이다. 복잡한 화학 물질을 한 줄로 표기할 수 있기 때문에 line notation (line-entry) 시스템으로 불리며 다음과 같은 몇 가지 특성을 갖는다[6]. 1) 각 원자는 해당 원자 기호로 표기한다. 2) 자유원자가(free valence)에 자동적으로 포화되는 수소는 표기를 생략한다. 3) 바로 옆에 표기되어 있는 원자는 서로 결합되어 있다. 4) 이중결합과 삼중결합은 각각 "="과 "#"로 표기한다. 5) 결합가지는 괄호(parentheses)로 표기한다. 6) 고리구조는 서로 연결되어 있는 원자에 숫자로 표시한다. 7) Tetrahedral carbon의 경우엔 '@'와 '@@'로 방향을 나타낼 수 있다. 4개의 결합을 가졌을 경우 나타나는 순서대로 왼쪽에서 오른쪽으로 판단한다. 첫 번째 결합(bond)의 시점에서 중앙 탄소를 바라보았을 때 반시계 방향이면 @, 시계 방향이면 @@로 표기한다.

2. 사용된 데이터 베이스

(1) Drugbank

Drugbank는 생물정보공학(Bioinformatics)과 화학정보공학(Cheminformatics) 분야에서 약물에 대한 다양한 정보를 제공하는 웹 기반 데이터베이스이다[1]. 총 14,528개의 약물을 제공하고 있으며, 2,358개의 FDA 승인 저분자의약품(Small Molecule Drug)를 포함한다.

본 연구에서는 Small Molecule, Biotech, Experimental, Nutraceutical, Illicit, Withdrawn의 6가지 분류기준을 따라 8가지로 세부 분류되는 14,528개의 전체 약물 중, SMILES에 대한 정보가 존재하는 11,281개의 Data1과 약물에 대한 2,841개의 ATC code까지 포함하는 Data2, 2가지로 나누어 진행한다.

(2) 한국한의약진흥원

한국한의약진흥원(National Institute for Korean Medicine Development, Nikom)은 다양한 한의약소재(한약재, 생약재, 약용식물)및 천연물에 대한 약리학적 정보를 제공한다. 본 연구에서는 한국한의약진흥원에서 제공하는 천연물 3가지를 SMILES 표기법으로 나타내어 모델에 적용하였다. 이를 통해 본 모델이 약물 재창출과 같은 맥락으로 유용하게 고려될 수 있음을 보이며, 기존에 제시되지 않은 식용 천연물, 화합물들의 약리학적 유효성을 검토하였다.

3. ATC

ATC code (Anatomical Therapeutic Chemical Classification System)는 의약품의 분류를 위해 사용되는 코드이다[6]. 세계보건기구(WHO) 산하 기관인 의약품 통계 방법을 위한 협력센터(WHO Collaborating Centre for Drug Statistics Methodology, WHOCC)가 이 코드를 관리한다. 약물을 그들이 작용하는 장기나 계통, 그리고 화학적 특성에 따라 영

표 1. ATC code 분류
Table 1. ATC code classification

Stage	Contents	Examples
1	The first letter designates one of the 14 major anatomical groups.	C-cardiovascular system
2	The therapeutic group is designated by two numbers.	C03-diuretic
3	Therapeutic/pharmacological subgroups are designated by one letter.	C03C-high ceiling diuretics
4	Chemical/pharmacological subgroups are designated by one letter.	C03CA-Sulfonamide diuretic
5	Chemicals are designated by two numbers.	C03CA01-Furosemide

문자와 숫자의 조합으로 7자리 고유코드를 부여한다. 각 단계의 ATC code는 약물의 작용부위에 따라 하나씩만 지정하기 때문에 한 약물이 여러 개의 ATC code를 가질 수 있다. 이 시스템에서 약물은 서로 다른 5단계에 따라 분류된다. 분류되는 단계와 그에 대한 내용을 표 1에 제시하였다.

본 연구에서는 Drugbank에서 제공하는 2,841개의 ATC code 데이터를 사용하며, ATC code가 4단계까지 같다면 기능적 유사도가 일치한다고 본다. 4단계는 약물의 치료적/약물학적 하위 그룹뿐만 아니라 화학적 그룹까지 가장 세밀하게 비교할 수 있기 때문이다.

4. 데이터 전처리

(1) SMILES 변환

원소기호와 숫자 그리고 특수문자(e.g. [ ], ( ), =, #, @)로 이루어져 원자간 결합형태를 나타내는 SMILES 표기법은 컴퓨터언어로 처리하면서 텍스트 자체로 인식되어야 하고, 오류의 발생을 방지하기 위해 원소기호와 겹치지 않는 조건에서 '[', ']', '(', ')'를 각각 'L', 'R', 'l', 'r'로 치환한다. 이는 알고리즘 입력 및 계산 시에 모두 적용한다.

(2) 다수의 ATC code 구분

Data2는 약물이름, SMILES, ATC code로 구성 되어있고, 한가지 약물이 여러 개의 ATC code를 가지는 경우에는 구분 기호(;)를 사용함으로써 컴퓨터가 여러 개로 인식할 수 있게 한다. Data2의 일부를 표 2로 제시하였다.

5. 예측 알고리즘

약물의 구조적 유사성과 기능적 유사성 사이의 관계를 분석하기위해 개별 약물의 SMILES를 입력으로 하여 data1의 11,281개 약물의 분자구조 유사도 점수를 계산한다. 이후, 상위 배치된 10개 약물을 각각의 점수와 함께 출력한다. 본 연구에서는 SMILES간 유사성을 수치화하기위해 4가지의 기계학습 알고리즘을 사용한다. 각 4개의 유사성 계산 알고리즘을, 두 약물 Serine ("N[C@@H]C(O)C(O)=O")과 Methionine ("CSCC[C@H](N)C(O)=O")의 SMILES를 예시로 두어 소개한다. 입력이 되는 SMILES는 간단히 변수 SMI_1, SMI_2로 표현한다.

(1) Edit distance

편집거리는 가장 널리 사용되는, 문자열 간 비교를 위한 계산방식 중 하나이다. 두 문자열(SMI_1, SMI_2)이 주어진 경우 이들 사이의 편집거리는 edit (SMI_1, SMI_2)로 표현하고, SMI_1를 SMI_2로 변환하는 데 필요한 '최소 편집 연산의 수'로 정의된다[8]. 편집연산은 문자열 요소에 대한 삽입, 삭제, 치환 세 가지만을 허용한다[9]. 앞서 제시한 예시 약물 Serine를 Methionine으로 변환하기 위해서는, 공통부분 "C(O)=O"를 제외한 앞쪽 문자열에 대해 2번의 치환("N"과"O"를 각각 "C"와"N"으로 치환), 4번의 삽입("SCC", "@"), 마지막으로 1번의 삭제("C") 연산이 필요하다. 결과적으로 두 입력약물에 대한 편집거리는, edit (Serine, Methionine)=7이다. 두 약물의 편집거리와 함께 $MAX(length(SMI_1), length(SMI_2)) = length(SMI_2) = 18$ 이므로, 구조적 유사도는 식 (1)과 같이 계산되며 유사도 점수는 $1-\frac{7}{18}\approx 0.61$이다.

$$Editsim\ (SMI_1,\ SMI_2) = 1-\frac{edit(SMI_1,\ SMI_2)}{MAX(length(SMI_1),\ length(SMI_2))} \quad (1)$$

(2) Normalized longest common subsequences

NLCS 알고리즘은 두 입력이 공통으로 갖는 문자열 요소의 최대 길이를 계산하여 유사도 점수를 산출하는 방식이다[10]. 두개의 약물 SMILES가 주어졌을 때, 공통 부분은 $LCS(SMI_1,\ SMI_2)$로 표현한다. 예시에 적용해보면, 공통으로 갖는 문자열 요소는 "C(O)=O"로, LCS(Serine, Methionine)=6이다.

$$NLCS(SMI_1,SMI_2)=\frac{length(LCS(SMI_1,SMI_2))^2}{length(SMI_1)\times length(SMI_2)} \quad (2)$$

식 (2)에 따른 두 약물의 표준화된 유사도 점수는 $\frac{6^2}{17\times 18}\approx 0.12$이다.

표 2. Data2의 ATC code 변환 예시
Table 2. ATC Converted Data2

Name	SMILES	ATC
Protriptyline	CNCCCC…	N06AA11
Methylergometrine	[H][C@@]...	G02AC01 G02AB01
Buclizine	CC(C)(C)C…	R06AE51 R06AE01
Chlorzoxazone	ClC1=CC…	M03BB03 M03BB53 M03BB73
Grepafloxacin	CC1CN(C…	J01MA11
Meprobamate	CCCC(C)(…	N05BC51 N05CX01 N05BC01

(3) LINGO similarity

LINGO Similarity 알고리즘은 문자열을 일정한 크기로 조각내 이를 다시 입체적으로 재구성하는 방법이다[11]. 이는 단순 분자간 유사성 계산에 사용될 뿐 아니라, 나아가 Quantitative Structure-Property Relationships 예측 모델에도 활용된다.

알고리즘을 통한 유사도 계산을 수행하기 이전에 'LINGO representation'과정이 반드시 선행되어야 한다. 'LINGO representation'은 입력 문자열을 일정한 길이 q를 갖는 하위 문자열들로 잘게 조각내 입체적으로 표현하는 방법이다. 일정한 길이 q로 조각난 하위 문자열들을 LINGOs라고 부르며, 중복되는 LINGOs가 존재하는 경우에 중복횟수만큼 빈도로 표시한다. 해당 알고리즘에서는 하위 문자열의 길이 q를 4로 고정한다[12]. 추가적으로, 결합 숫자가 포함된 SMILES (e.g. Phentermine: CC(C)(N)CC1=CC=CC=C1)의 경우에는 숫자를 0으로 치환한 후 representation을 진행한다. 예시 약물 Serine, Methionine에 대한 적용 결과는 표 3에 제시하였다.

LINGO representation과정 이후, 두 약물의 하위 문자열(LINGOs)의 총 가짓수인 m을 구한다. 이 때 중복을 허용하지 않으며, 빈도에 관계없이 총 가짓수만을 고려한다. 표 3을 참고할 때, 두 약물의 LINGOs는 각각 13, 15이고, 이들은 "C(O)", ")C(O", "(O)=", "O)=O" 4가지의 LINGOs를 중복으로 가지므로 총 가짓수 m은 13+15-4=24이다.

$$LINGOsim = \frac{\sum_{i=1}^{m} 1 - \frac{|N_{SMI_1,i} - N_{SMI_2,i}|}{N_{SMI,i} + N_{SMI,i}}}{m} \quad (3)$$

표 3. 2가지 약물의 LINGO Representation
Table 3. LINGO Representation of 2 drugs

Serine	Frequency	Methionine	Frequency
N[C@	1	CSCC	1
[C@@	1	SCC[	1
C@@H	1	CC[C	1
@@H]	1	C[C@	1
@H]C	1	[C@H	1
H]C(	1	C@H]	1
]C(O	1	@H](	1
C(O)	2	H](N	1
(O)C	1	](N)	1
O)C(	1	(N)C	1
)C(O	1	N)C(	1
(O)=	1	)C(O	1
O)=O	1	C(O)	1
		(O)=	1
		O)=O	1

m이 결정되면 모든 LINGOs에 1부터 m의 인덱스를 부여한다. 식 (3)과 같은 유사도 계산과정에서 분자의 $N_{SMI_1,i}$항목은 전체 LINGOs 중 i의 인덱스를 갖는 LINGOs의 SMI1 빈도에 해당하며, 두 약물에 대하여 중복되지 않는, 즉 한쪽 약물에만 존재하는 LINGOs는 0으로 계산되어 유사도 점수에 영향을 미치지 못한다. 이와 다르게 두 약물의 LINGO representation에 공통으로 존재하는 LINGOs의 경우에는 0이 아닌 값으로 계산되고, 서로 다른 빈도에 따라 유사도 점수에 기여한다. 이를 바탕으로 계산된 두 약물 Serine, Methionine의 유사도 $LINGOsim(Serine, Methionine)$이다.

(4) LINGO cosine

LINGO cosine 알고리즘은 LINGO based TF cosine similarity라고도 불리며, 해당 알고리즘에서 파악된 LINGOs의 수를 나타내는 빈도를 가중치로 변환하는 과정이 추가된다[12]. 이는 특정 시퀀스가 과도하게 많이 반복되는 형태(e.g. 빈도=100)의 고분자 화합물과의 유사도 점수를 계산할 때 발생하는 오류를 해결할 수 있다.

$$TFweight_{L,S} = \begin{cases} 1+\log_{10}(termfrequency_{L,S}), & if\ termfrequency_{L,S}>0 \\ 0, & otherwise \end{cases}$$

예를 들어 특정 LINGOs에 대한 빈도가 2인 경우에는 $1 + \log_{10}(2) \approx 1.3$로, 빈도가 100인 고분자 화합물의 경우 이를 $1 + \log_{10}(100) = 3$으로 계산하여 가중치로 변환한다.

$$CosSim(SMI_1, SMI_2) = \frac{\sum_{i=1}^{m} V_{SMI1,i} V_{SMI2,i}}{|V_{SMI1}||V_{SMI2}|} \quad (4)$$

두 약물의 가중치 변환을 m개의 전체 LINGOs에 대하여 실시한다. 결과적으로 두 약물 모두 m개의 가중치를 가지는데, 이를 각각의 약물에 대해 길이가 m인 두 벡터($V_{SMI1,i}V_{SMI2,i}$)로 표현한다[13]. $V_{SMI1,i}V_{SMI2,i}$는 각 약물에 대한 '특성벡터'이며, 최종 유사성은 식 4와 같이 두 벡터 사이의 코사인 각도에 의해 결정된다.

두 약물 Serine, Methionine에 대한 특성벡터는 표 3를 참고하여 구할 수 있고, 두 벡터는 각각 V_{Serine} = [1,1, 1,1,1,1,1,1.3,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0], $V_{Methionine}$ = [0,0,0,0,0,0,0,1,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1] 이다. 최종적으로 유사도 $CosSim(Serine, Methionine)$는 0.30이다

6. 실험

(1) 알고리즘 선정1

4가지 알고리즘(Edit distance, NLCS, LINGO_sim, LINGO_cos)을 구현하고 한가지의 고정된 약물의 분자구조를

입력하여 해당 약물과 분자구조 유사성이 높은 10개의 약물을 산출한다. 이를 통해 1차적 성능평가를 진행하고, 각 모델이 갖는 한계를 도출한다.

(2) 알고리즘 선정2

알고리즘 선정2는 data2와 알고리즘 선정1에서 선정된 두 가지 후보모델(Edit distance, LINGO cosine)을 사용한다. 1차적으로 한가지 약물에 대한 모든 약물의 구조적 유사도 점수를 계산한다. 계산된 유사도 점수를 바탕으로 4개의 그룹으로 나누는데 유사도 점수 0.6~0.7은 group 1, 0.7~0.8은 group 2, 0.8~0.9는 group 3, 0.9~1.0은 group 4로 분류된다.

2차적으로 각 그룹별로 입력 약물과 ATC code가 4단계 이상으로 일치하는 약물의 수를 센다. 이때, 한 약물에 대한 ATC code가 2개 이상인 경우, 한가지만 일치하여도 그 수를 세며, 2개의 ATC code 모두 4단계 이상 일치하는 경우에는 2로 센다. 실험 결과는 한가지 입력 약물에 대해, 2개의 모델과 4개 그룹에 따라 분류되어 2*4의 배열 형태로 나타난다.

(3) LINGO cosine 알고리즘

알고리즘 선정2에서 도출된 모델이 개별 약물에 대해 어떻게 활용가능한지 ATC code를 활용하여 직접 검증한다. 기준 약물의 분자구조를 입력으로 하여 해당 약물과 유사도가 높은 상위 10개의 약물을 배치하고 ATC code를 기준으로 분자구조-기능쌍을 나타낸다.

III. 연구 결과 및 고찰

1. 알고리즘 선정1

4가지 알고리즘이 기준 약물의 SMILES code를 입력으로 받을 때 Data1에서 구조적 유사도를 하위 정렬하는 과정을 살펴보았다. 결과를 바탕으로 각각이 구조적 유사도를 공정하게 산출할 수 있는지 검토한다. 그림 1A에서 Edit distance 모델에 따른 결과를 살펴보면, 상위 10개의 약물이 다른 유사도 값을 가진다. 따라서 해당 모델의 검증을 통해 유사도-기능 쌍의 경향성을 파악할 수 있을 것이라 판단하였다.

그림 1B에서 NLCS 모델에 따른 결과를 살펴보면, 상위 10개의 약물 모두 같은 최대값(29)을 가진다. 따라서 패턴 약물과 유사성을 비교하고 구별하는데 어려움이 있다. SMILES는 약물의 결합 순서에 따른 정보도 담고 있지만, NLCS 모델은 알고리즘 자체적으로 문자열의 연속을 고려하지 않고 공통 부분의 최장길이 수열을 검색하는 것이 중복 최대값의 원인으로 사료된다.

그림 1C에서 LINGO Similarity 모델에 따른 결과를 살펴보면, 상위 10개의 약물 모두 데이터상의 최장문자열(SMILES)임을 알 수 있다. 즉, 고분자의 약물들이 약물 입력과 관계없이 동일하게 산출되는 결과를 보였다. 가장 짧은 SMILES를 가진 약물은 N-Valeric Acid로 길이가 10인 반면, 가장 긴 SMILES를 가진 약물 Inulin은 길이가 1,526으로, 고분자 약물에서 저분자 약물에 비해 SMILES 길이가 최대 152.6배 길다. 따라서, 빈도수를 온전히 유사도에 반영하는 해당 모델에서는 고분자 약물이 높은 유사도 값을 가질 수밖에 없으므로 채택할 수 없다고 판단하였다.

그림 1D에서 LINGO cosine 모델에 따른 결과를 살펴보면, 상위 10개의 약물이 다른 유사도 값을 가지며, 고분자 또는 저분자의 경향성이 편향되어 있지 않았다. 따라서 해당 모델의 검증을 통해 유사도-기능 쌍의 경향성을 파악할 수 있을 것이라 판단하였다. 최종적으로, 알고리즘 선정1을 통해 두 가지 알고리즘(Edit_dis, Lingo_cos)을 채택한다.

2. 알고리즘 선정2

표 4는 알고리즘 선정2에 관한 결과 중 일부를 나타낸 것으로, Group 항목은 SMILES유사도 점수에 따른 분류이다. Edit_dis와 LINGO_cos항목은 각각 Edit distance, LINGO cosine 알고리즘에 기반한 두 모델이 해당 그룹 내에서 갖는 기준 약물과 ATC code가 일치하는 약물의 수를 의미한다. 이에 따라 한가지 약물은 그룹에 따라 4개의 행을 가진다.

4가지 모델 중, 알고리즘 선정1에서 선별된 2가지 모델(edit distance, LINGO cos)이 각각 구조적 유사성을 기능적 유사성에 얼마나 연관 지을 수 있는지에 대한 2차 검증을 실시하였다. 해당 과정에서 그룹명은 숫자가 높을수록 해당 모델이 구조적으로 매우 유사하다고 판단했음을 의미한다.

그림 2에서 Edit distance 모델은 group 1을 제외하고는 모든 그룹에 대해서 ATC code 일치약물의 수가 0으로, 모델이 구조적으로 높은 유사도를 보인다고 판단한 두 약물에 대해서 기능적 유사성을 발견할 수 없다. 반면에 LINGO cosine 모델의 결과는 더 높은 구조적 유사성을 갖는 것으로 판단한 그룹에서 실제로 더 많은 기능적 유사성을 보였다. 이를 통해 Edit distance 알고리즘을 적용한 모델이 SMILES를 입력으로 한 구조-기능 유사도 모델에는 부적합하다고 판단하였다. 그림 2에 나타난 LINGO cosine 모델은 group 4에서 기능적 유사도가 가장 높았으며, '구조가 유사한 약물은 비슷한 기능을 할 것이다'라는 가설을 상당부분 입증하였다. group 2에서 3구간으로 진행되면서 일치 약물 평균이 떨어지는 경향을 보인다. ATC code 데이터가 충분히 확보되지 않았으므로 11,281개의 약물에 ATC code가 1:1 대응하지 않는다. 또한 ATC code를 1~2개 갖는 약물도 있는 반면에, 특정 약물은 ATC code를 30개 이상 보유하기 때문에 표준편차가 다소 크다는 한계에서 비롯되는 것으로 보인다. 그래프의 해당 부분을 제외한 구간에 대해서는 선형적으로 증가하는 경향을 나타낸다.

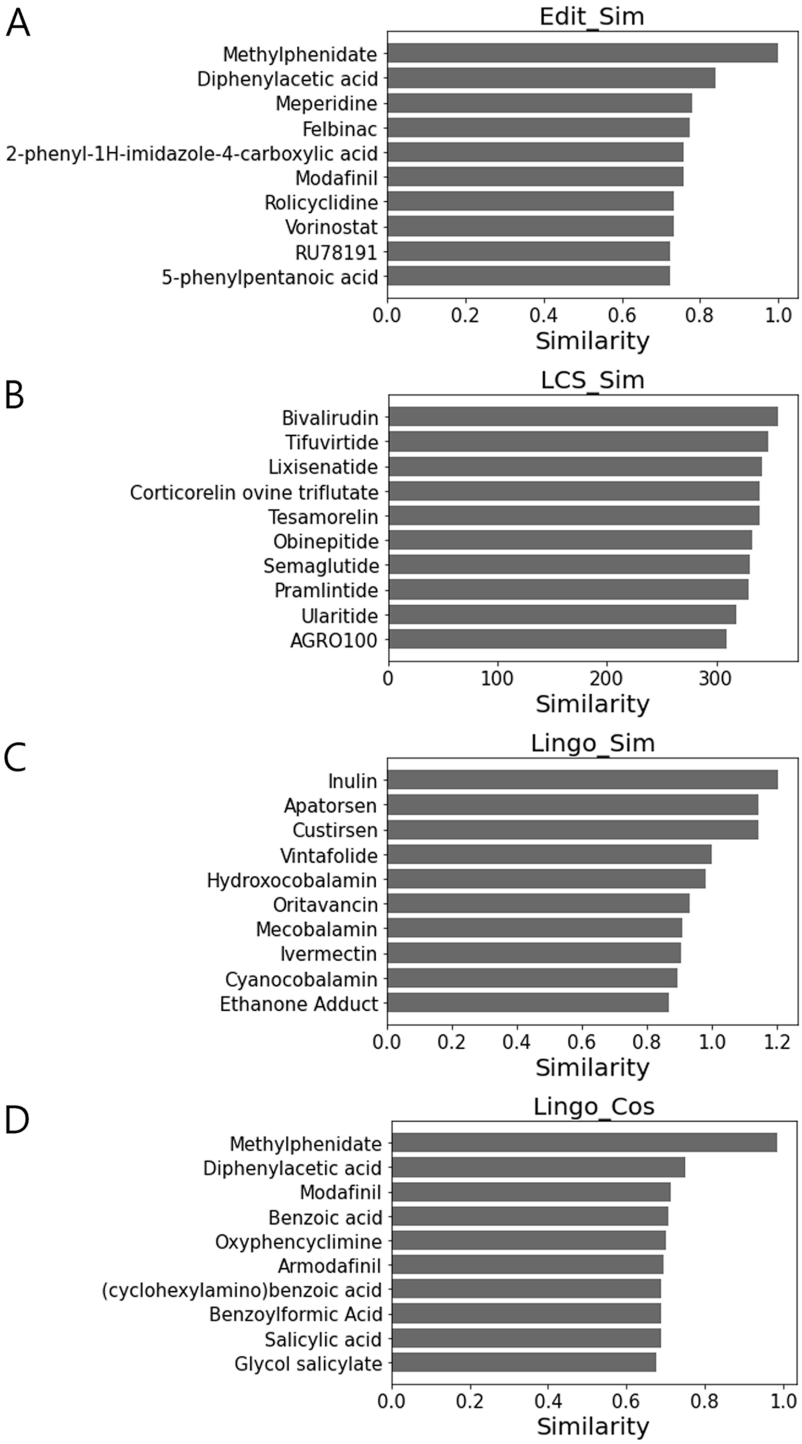


그림 1. SMILES code의 알고리즘별 유사도 평가(A. 편집거리, B. 문자열 최대길이, C. LINGO 유사성, D. LINGO-코사인 알고리즘)
Fig. 1. Evaluating the similarity of SMILES codes by algorithms (A. Edit distance, B. Longest common subsequences, C. LINGO similarity, D. LINGO cosine algorithm)

또한 제시된 값을 전구간에 대해 계산된 ATC code의 평균값으로 나눈 데이터의 백분율을 나타낸 값이다. Edit distance는 전반적으로 유사도 0.6 이상에서 38.95%의 기능적 유사도를 예측하나, 구조 유사도가 높은 group 2부터 4에서는 기능적 유사도를 전혀 감지하지 못한다. 반면 LINGO cos는 유사도 0.6 이상에서 82.59%의 기능적 유사도를 예측하였다. 또한 group 4에 가까워질수록 더 높은 일치율을 보인다.

표 4. 약물에 대한 그룹화 예시
Table 4. Examples of grouping for drugs

SMILES	ATC	Group	Edit	LINGO
CSCC[C...	V03AB26	1	1	0
CSCC[C...	V03AB26	2	0	0
CSCC[C...	V03AB26	3	0	0
CSCC[C...	V03AB26	4	0	1
C[C@@H…	L01XX10	1	1	0
C[C@@H…	L01XX10	2	0	0
C[C@@H…	L01XX10	3	0	0
C[C@@H…	L01XX10	4	0	0
CCC(O)(...	N05CM08	1	1	0
CCC(O)(...	N05CM08	2	0	0
CCC(O)(...	N05CM08	3	0	0
CCC(O)(...	N05CM08	4	0	1

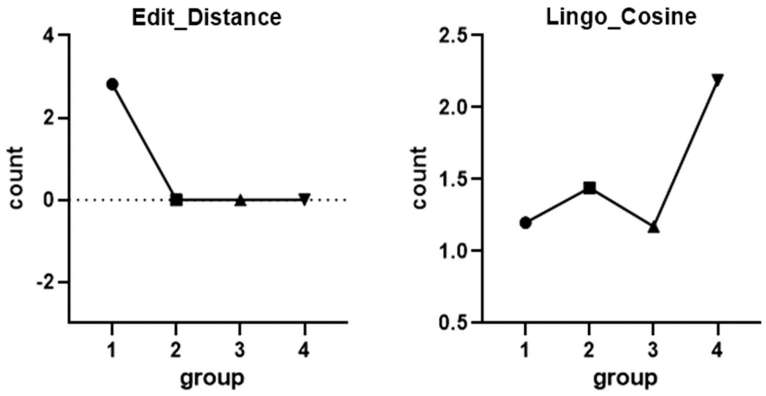


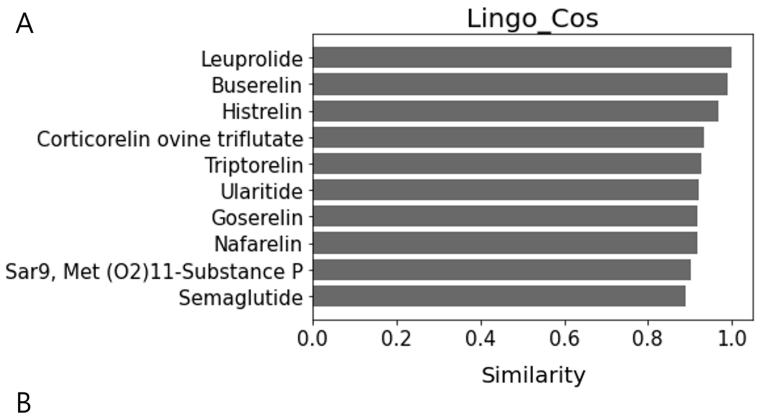
그림 2. 그룹별 알고리즘 세부 유사도 평가
Fig. 2. Evaluating of the detailed similarity of algorithms by group

3. LINGO cosine 알고리즘

그림 3은 LINGO cosine 모델에서 입력 약물 Leuprolide에 대한 분자구조 유사도 상위 10개의 약물을 나타낸다. 그림 3B는 그림 3A의 결과와 ATC code 데이터를 결합하여 나타낸다. ATC code 데이터가 존재하지 않는 약물은 None으로 표기하였다.

LINGO Cosine모델을 통해서 Input약물 Leuprolide에 대해 분자구조가 유사한 상위 10개의 약물을 나타냈다. 이를 통해 그림 3A에서 ATC code를 기준으로 약물 분자구조 유사도-기능 쌍의 표를 제시하였다. 상위 8번째 약물 Nafarelin과 10번약물 Abarelix를 제외한 약물들은 기준 약물 Leuprolide에 대해 ATC code가 4단계(L02AE)까지 동일하다. 따라서, 이들은 동일한 치료적/약리학적/화학적 기능을 한다고 판단할 수 있다. 10번 약물 Abarelix는 ATC code가 2단계(L02)까지 동일하다. 2단계는 같은 치료적 그룹을 지정하므로(e.g. 내분비계통 치료제) 10번 약물은 기준약물과 동일한 약리학적/화학적 기능을 하지 않지만, 동일한 치료적 기능을 한다고 판단할 수 있다. 기준약물 Leuprolide (L02AE02)과 Nafarelin (H01CA02)는 전혀 다른 ATC code가 부여되어 있으므로 범용적으로 사용되는 영역은 다르다고 여겨진다. 그러나 Leuprolide와 Nafarelin은 체외 수정주기에서 성선 자극 호르몬 방출 호르몬 유사체의 선택과 관련된 결과에 차이가 있는지 확인하는 연구[14] 에서 같은 기능에 대해 비교하는 직접적인 대조군으로 활용된 바 있다. 위와 같은 결과에서 모델을 통해 입력한 약물과 높은 분자구조 유사도를 보이는 약물들은 같은 기능을 공유하는 것을 보았다. 또한 유사도가 충분히 높다면 그것이 범용적으로 사용되는 분야(ATC code)는 다르더라도 대조군으로 활용된 연구를 통해 비슷한 약물학적 기능을 할

A



B

Rank	Drug	Similarity	ATC code
1	Leuprolide	1	L02AE02
2	Buserelin	0.989296	L02AE01
3	Histrelin	0.967759	L02AE05
4	Corticorelin ovine triflutate	0.933443	None
5	Triptorelin	0.926430	L02AE04
6	Ularitide	0.921192	None
7	Goserelin	0.918526	L02AE03
8	Nafarelin	0.917371	H01CA02
9	Sar9, Met(O2)11-Substance P	0.903120	None
10	Abarelix	0.889287	L02BX01

그림 3. Leuprolide 계열의 유사도 평가 및 구조-유사도 비교
Fig. 3. Similarity assessment of Leuprolide Series and molecular structural-similarity comparison with LINGO cosine

수 있다는 가능성을 제시할 수 있다.

4. 천연물을 통한 검증

앞선 모델에 의해 분자구조의 유사성이 기능적 유사성을 평가하는데 활용될 수 있는 가능성을 보았다. 따라서 유사도가 충분히 높다면 약리학적 측면에서 기존에 존재하던 식용 천연물, 화합물의 새로운 기능을 제시할 가능성이 있다. 본 연구에서는 이러한 가능성을 보이고자 한국한의약진흥원에서 제공하는 천연물 4가지를 SMILES 표기법으로 나타내어 해당 모델에 적용하였다.

천연물 결명자의 순도 96% 성분의 SMILES를 입력으로 하여 본 모델에 적용한 결과를 Result의 그림 4A에 제시하였다. 배치되는 약물 중 Acarbose를 제외한 약물의 ATC code 데이터가 존재하지 않아 실험논문을 통해 약물 및 천연물의 기능을 평가하였다. 또한 약물로서 기능을 하지 않는 단일 분자구조의 데이터는 제외하였다. 결명자의 풍부한 수용성 섬유소는 생리 활성을 나타내는 기존의 수용성 섬유소와 유사한 이화학적 특성을 나타내어 혈중 지질 저하 및 혈당 상승 억제 등의 기능성 소재임이 알려져 있다[15]. 그림 4A에서 검증대상이 아닌 약물 중, 결명자와 가장 상위 유사도를 갖는 Ginsenoside Rb1은 비만치료제, 고혈당치료제, 당뇨병 약으로서 사용되는 천연물 성분이다[16]. Acarbose 또한 인슐린 분비를 촉진하여 혈당을 억제하는 약물로 널리 알려져 있다[17]. 반면 Echinacoside는 MPTP로 유도된 Parkinson's diesase의 쥐 모델에서 신경보호효과를 보였다[18]. 다음과 같은 근거로 결명자의 신경보호효과 기능을 찾아본 결과, 알코올에 의한 기억력 손상 및 fEPSP의 감소를 억제한다는 사실을 확인할 수 있었다[19]. 즉, 모델을 통해 유추된 기능이 실제 진행된 연구를 통해 검증되었음을 확인하였다.

Result의 그림 4.B에는 천연물 매생이의 순도96% 성분의 SMILES를 입력으로 하여 본 모델에 적용한 결과를 제시하였다. Ursodeoxycholic acid를 제외한 약물의 ATC code가 제공되지 않아 실험 논문을 통해 기능을 평가하였고, 약물로서 기능을 하지 않는 단일 분자구조의 데이터 역시 제외하였다. 매생이는 많은 당질과 단백질을 함유하고 있으며 쥐 모델에서 혈청 지질개선 효과 및 체내 총 콜레스테롤 함량을 낮추는 효과를 보인다[20]. 검증 대상 약물 중, Ursodeoxycholic acid는 콜레스테롤로 인한 담석증의 치료에 효과적인 약물로

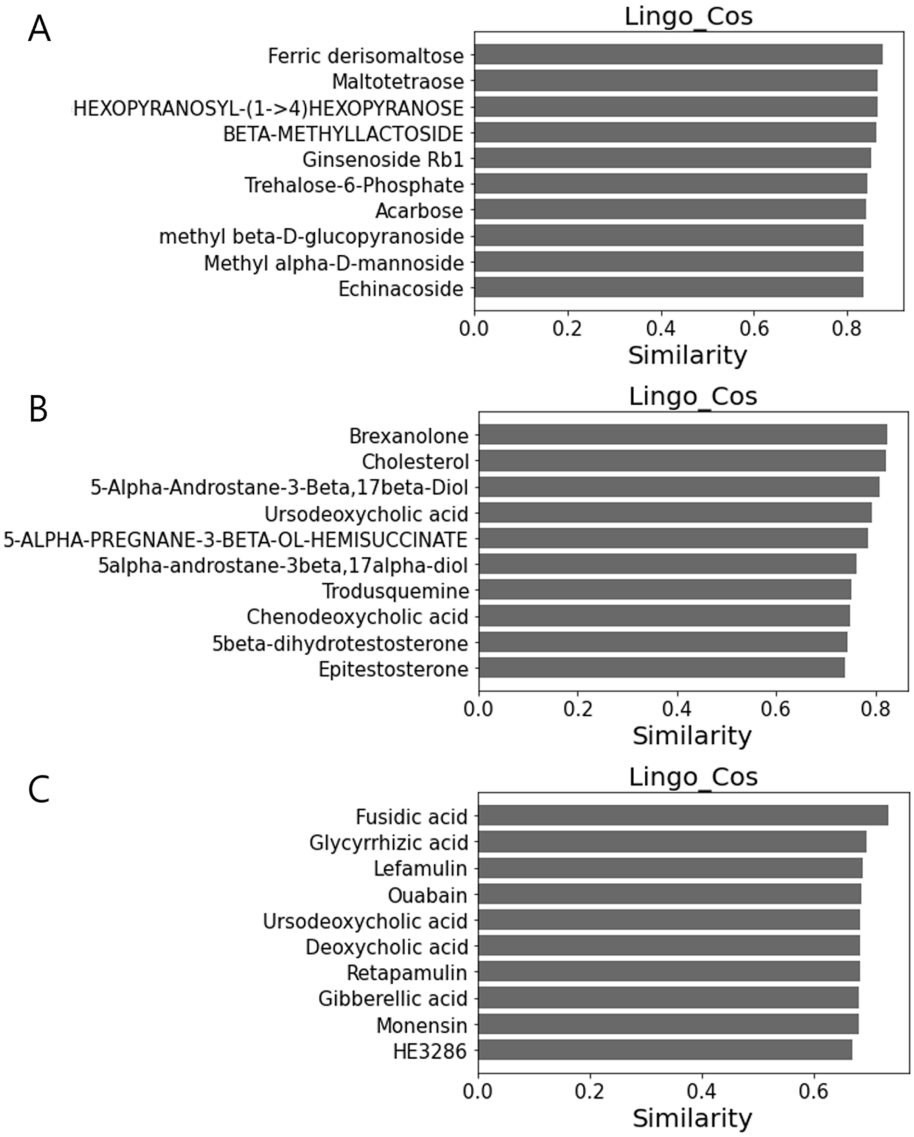


그림 4. 천연물의 유사도 평가 및 구조-유사도 비교(A. 결명자, B. 매생이, C. 찔레꽃)
Fig. 4. Similarity assessment of natural products and its similarity exclusive of Single Molecular Structure comparison with LINGO cosine (A. Cassia Seed, B. Capsosiphon fulvescens, C. Rosa multiflora)

평가된다[21]. Trodusquemine는 부작용 없이 콜레스테롤 담석증을 치료하여 Atherosclerosis(죽상동맥경화증) 감소에 효과적인 약물로 평가되었다[22]. 약물 Chenodeoxycholic acid 또한 콜레스테롤 담석증을 6개월 내 효과적으로 제거하여 의학적 치료로 활용될 수 있음을 보였다[23].

반면 매생이의 분자구조와 가장 높은 유사성을 갖는 Brexanolone는 산후 우울증 치료에 효과적인 약물로서 미 FDA 최초로 여성용 산후 우울증 치료 약물로서 승인되었다[24]. 이와 관련한 매생이의 실험연구는 진행되어 있지 않다. 그러나 앞선 유사도 검색식의 검증에서 도출된 결과를 통해 '매생이가 산후 우울증 치료에 효과적일 수 있다'는 실험연구의 가설로 설정할 수 있는 가능성에 의미를 갖는다.

Result의 그림 4C을 통해 찔레의 순도 99% 주성분의 SMILES를 본 모델에 적용한 결과를 제시하였다. 배치되는 약물 중 Fusidic acid를 제외한 약물의 ATC code에 관한 데이터가 존재하지 않아 실험 논문을 통해 기능을 평가하였다. 천연물 찔레는 메티실린에 내성이 있는 포도상구균에 대한 항균작용을 나타낸다는 연구가 보고된 바 있다[25]. 검증대상 약물 중, Fusidic acid는 피부의 세균감염으로 인한 2차감염을 예방하는 항생제, 연고 '후시딘'의 주성분 퓨시드산 나트륨(Sodium fusidate) 계열의 약물이다. Retapamulin은 피부감염에 대한 국소 항감염제로써 사용된다[26]. 두 약물은 공통적으로 항균, 항감염작용을 하며, 무피로신 및 메티실린과 같은 국소항생제에 내성을 갖는 포도상구균에 대한 *in vitro*

항균활동이 연구된 바 있다[27]. Glycyrrhizic acid는 감초의 뿌리 추출물로, 체내 작용으로는 낮은 복용량에 대해서 항염증, 항균, 항바이러스, 항종양 등의 광범위한 약리학적 효과를 나타낸다. 과거 2005년 SARS virus에 대해 항바이러스성 효능이 검증된 바 있다[28]. 이에 더하여, 찔레와 가장 높은 구조유사성을 갖는 Fusidic acid는 200일간 처리하여 진행한 실험에서 당뇨수준이 실험군은 52%, 반면 대조군 71%로 당뇨유발이 억제됨을 확인한 바 있다[25]. 같은 맥락으로, 찔레는 sulfonylureas와 같은 작용기전으로, 랑게르한섬의 베타세포를 자극하여 인슐린의 분비를 증가시키는 작용을 한다. 그 결과, 당뇨가 유도된 쥐 모델에 대해 혈당치료제인 glibenclamide 와 동등하지는 않으나 상당한 혈당저하 수준을 보였다[29]. 이처럼 찔레가 가장 유사도상 가장 상위 배치된 약물과만 특이적으로 기능을 공유한다는 결과는 앞선 매생이에 대한 유사도 검증과정에서 언급한 '천연물의 용도에 대한 새로운 가설 제시 가능성'을 뒷받침하는 것으로 볼 수 있다. 천연물을 신약개발에 사용하는 것은 식물의 성장과정 중 다양한 미생물로부터 스스로를 보호하는 특성을 이용하는 것이며, 이를 통해 충분한 안정성을 확보할 수 있다. 이에 따라 본 모델을 통해 기존약물의 기능을 천연물에서 찾아내는 과정은 신약개발과정에 유용하게 고려될 수 있다.

IV. 결 론

본 연구에서는 텍스트 기반 유사도 검색 알고리즘을 선별하고 Drugbank의 SMILES 데이터에서 약물-효능 관계를 식별하였다. 제안된 검색 모델에 따라 높은 분자구조 유사도를 갖는 약물에서 기능적 유사도가 증가하는 것을 ATC code를 통해 보였다. 또한 제안된 모델이 어떻게 활용될 수 있는지 개별 약물과 천연물을 통해 제시하였다.

분자구조 이미지를 SMILES로 변환하는 인공신경망 알고리즘에 대한 연구가 지속적으로 이루어지고 있으나[2,3] 입체화학(stereochemistry)적 측면에서는 정체되어 있는 추세이며, 본 연구에서 활용된 SMILES 또한 입체화학 정보를 담고 있지 않다. 따라서 구현된 모델은 원자의 공간 배열만 다른 입체 이성질체(Stereoisomer)를 같은 약물로 판단하게 된다는 데에 한계가 있다. 또한 11,886개의 약물과 그에 비해 비교적 적은 2,841개의 ATC code를 활용하였다. 제시한 모델이 보다 높은 정확성과 안정성을 갖기 위해서는 ATC code와 약물이 최소 1:1 대응을 갖도록 추가되어야 한다. 그리고 본 모델은 약물의 분자구조를 문자열로 변환하여 평가하므로, Inulin과 같이 SMILES로 변환하였을 때 1200자가 넘는 고분자 약물은 전 데이터에서 구조적 유사도를 검색하는데 많은 시간이 소요된다. 해당 약물이 input 약물로 입력될 경우 상당수의 약물과 구조적 유사성을 중복하여 공유하므로 기능적 유사도의 검증에서 높은 정확도를 얻는데 어려움이 있다. 선택한 알고리즘은 해당 문제에서 비교적 높은 정확도를 가지나, 이 또한 해당 문제에서 완전히 벗어나지 못했다. 이러한 한계를 극복하고 데이터가 확장된다면 약물 개인별로 데이터를 학습하고 예측하는 ML/DL에서도 성과를 보일 것이며 보다 다양한 약물을 제시할 수 있을 것이다.

본 연구에서 제시한 모델은 모든 물질을 구성하는 분자구조 유사도 검색 모델이기 때문에 추가적인 데이터와 결합이 용이하다. 따라서 향후 연구에서는 본 모델에 ATC code 데이터 외에 다양한 데이터를 결합하여 연구범위를 확장하고 예측가능성을 향상시키는 등의 긍정적 효과를 기대할 수 있다. 예를 들어, 약물에 대한 부작용 데이터를 결합하면, 약물의 치료적 기능 뿐만 아니라 부작용을 추가적으로 예측할 수 있다. 더불어 개인별로 약물 효능에 편차가 발생하는 원인인 유전자 반응 데이터를 결합하면 예측가능성을 높일 수 있다.

본 연구에서 제시한 모델은 알려진 약물의 정보를 이용하여 기존 약물 또는 천연물의 새로운 기능을 발견할 수 있는 '약물 재창출'의 기반이 될 것이라 기대된다.

References

[1] Weininger, D. Smiles, a Chemical Language and Information-System .1. Introduction to Methodology and Encoding Rules. Journal of Chemical Information and Computer Sciences. 1988;28(1):31-36.

[2] Rajan K, Zielesny A, Steinbeck C. DECIMER: towards deep learning for chemical image recognition. J Cheminformatics. 2020;1291:1-9.

[3] Staker J, Marshall K, Abel R, McQuaw C. Molecular Structure Extraction from Documents Using Deep Learning. Journal of Chemical Information and Modeling. 2019;59(3):1017-1029.

[4] Oprea T, Mestres J. Drug repurposing: far beyond new targets for old drugs. The AAPS journal. 2012;14(4):759-763.

[5] World Health Organization. WHO collaborating centre for drug statistics methodology. ATC/DDD index 2011. World Health Organization2011WHO Collaborating Centre for Drug Statistics Methodology. ATC/DDD index. 2011.

[6] Gasteiger J, Engel T. Chemoinformatics: a textbook; John Wiley & Sons. 2006.

[7] Law V, Knox C, Djoumbou Y, Jewison T, Guo A, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Research. 2014;42(D1):D1091-1097.

[8] Marzal A, Vidal E. Computation of Normalized Edit Distance and Applications. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1993;15(9):926-932.

[9] Levenshtein V. Binary codes capable of correcting deletions, insertions, and reversals. InProceedings of the Soviet physics doklady. 1966;707-710.

[10] Kumar S, Rangan C. A Linear-Space Algorithm for the Lcs Problem. Acta Informatica. 1987;24(3):353-362.

[11] Vidal D, Thormann M, Pons M. LINGO, an efficient holo-

graphic text based method to calculate biophysical properties and intermolecular similarities. Journal of Chemical Information and modeling. 2005;45(2):386-393.

[12] Öztürk H, Ozkirimli E, Ozgur A. A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction. BMC Bioinformatics. 2016;17(1):1-11.

[13] Bilenko M, Mooney R. Adaptive duplicate detection using learnable string similarity measures. In Proceedings of the Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003;39-48.

[14] Martin M, Givens C, Schriock E, Glass R, Dandekar P. The choice of a gonadotropin-releasing hormone analog influences outcome of in vitro fertilization treatment. American journal of obstetrics and gynecology. 1994;170(6);1629-1634; discussion 1632-1624.

[15] Hong KH, Choi WH, Ahn JY, Jung CH, Ha TY. Physicochemical properties of ethanol extracts and dietary fiber from Cassia tora L. seed. The Korean Journal of Food and Nutrition. 2012;25(3):612-619.

[16] Zhou P, Xie W, He S, Sun Y, Meng X, Sun G, Sun X. Ginsenoside Rb1 as an Anti-Diabetic Agent and Its Underlying Mechanism Analysis. Cells. 2019;8(3):204.

[17] Chiasson J, Josse R, Leiter L, Mihic M, Nathan D, Palmason C, Cohen R, Wolever T. The effect of acarbose on insulin sensitivity in subjects with impaired glucose tolerance. Diabetes Care. 1996;19(11):1190-1193.

[18] Zhao Q, Gao J, Li W, Cai D. Neurotrophic and neurorescue effects of Echinacoside in the subacute MPTP mouse model of Parkinson's disease. Brain research. 2010;1346:224-236.

[19] Kwon H, Cho E, Jeon J, Lee Y, Kim D. Effect of an Ethanol Extract of Cassia obtusifolia Seeds on Alcohol-induced Memory Impairment. Journal of Life Science. 2019;29(5):564-569.

[20] Kwon MJ, Nam TJ. Effects of Mesangi (Capsosiphon fulvecens) powder on lipid metabolism in high cholesterol fed rats. Journal of the Korean Society of Food Science and Nutrition. 2006;35(5):530-535.

[21] Bachrach W, Hofmann A. Ursodeoxycholic acid in the treatment of cholesterol cholelithiasis. Digestive diseases and sciences. 1982;27(8):737-761.

[22] Thompson D, Morrice N, Grant L, Le Sommer S, Lees EK, Mody N, Wilson H, Delibegovic M. Pharmacological inhibition of protein tyrosine phosphatase 1B protects against atherosclerotic plaque formation in the LDLR-/-mouse model of atherosclerosis. Clinical Science. 2017;131(20):2489-2501.

[23] Danzinger R, Hofmann A, Schoenfield L, Thistle, J. Dissolution of cholesterol gallstones by chenodeoxycholic acid. New England Journal of Medicine. 1972;286(1):1-8.

[24] Kanes S, Colquhoun H, Gunduz-Bruce H, Raines S, Arnold R, Schacterle A, Doherty J, Epperson C, Deligiannidis K, Riesenberg R, et al. Brexanolone (SAGE-547 injection) in post-partum depression: a randomised controlled trial. Lancet. 2017; 390(10093):480-489.

[25] Eslami G, Taheri S, Ayatollahi S, Malek G, Pourkaveh B. Comparison of Rosa Nutkana Sepal Extract with Synthetic Antibiotics for Treatment of Methicillin Resistant Staphylococcus Aureus Isolated from Patients with Sty. Archives of Clinical Infectious Diseases. 2011;6(suppl):7-11.

[26] Odou M, Muller C, Calvet L, Dubreuil L. In vitro activity against anaerobes of retapamulin, a new topical antibiotic for treatment of skin infections. Journal of antimicrobial chemotherapy. 2007;59(4):646-651.

[27] Park SH, Kim JK, Park K. In Vitro Antimicrobial Activities of Fusidic Acid and Retapamulin against Mupirocin- and Methicillin-Resistant Staphylococcus aureus. Annals of Dermatology. 2015;27(5):551-556.

[28] Hoever G, Baltina L, Michaelis M, Kondratenko R, Baltina L, Tolstikov G, Doerr H, Cinatl J. Jr. Antiviral activity of glycyrrhizic acid derivatives against SARS-coronavirus. Journal of medical chemistry. 2005;48(4):1256-1259.

[29] Ivorra M, Paya M, Villar A. Hypoglycemic and insulin release effects of tormentic acid: a new hypoglycemic natural product. Planta Medica. 1988;54(4):282-285.