신약 발견을 위한 top-K 검색 엔진의 개발

서인*, 이승민**, 무하메드 이자즈 아메드*, 채송이*
*포항공과대학교 창의 IT 융합공학과
**포항공과대학교 컴퓨터공학과
e-mail : iseo@dblab.postech.ac.kr, smlee@dblab.postech.ac.kr, ejaz629@gmail.com, sychae@dblab.postech.ac.kr

Development of a top-K search engine for drug discovery

In Seo*, Seungmin Lee**, Muhammad Ejaz Ahmed *, Songyi Chae*
*Dept. of Creative IT Engineering, POSTECH
**Dept. of Computer Science and Engineering, POSTECH

요 약

신약 개발은 고부가가치를 창출하는 차세대 전략 산업으로 주목받고 있지만, 동물 실험과 임상 시험에 막대한 비용이 필요한 고위험-초고소득(high risk-super high return) 산업이다. 따라서 신약 후보군의 선정이 매우 중요하며 약물 유사도를 랭킹함수를 사용하는 top-k 질의 처리를 통해 후보군을 효과적으로 선정할 수 있다. 본 논문에서는 ChEMBL 데이터베이스[4]에 존재하는 화합물들 중 사용자가 원하는 특성을 갖는 k 개의 화합물들을 후보군으로 추천해주는 검색 엔진을 개발하였다.

1. 서론

고령화 사회로의 진입으로 신약 개발은 고부가가치를 창출하는 차세대 전략 산업으로 주목받고 있다. 신약 발견 과정은 고위험-초고소득(high risk-super high return) 산업인 신약 개발의 비용 절감에 중요한 역할을 한다. 신약 개발 절차 중 동물 실험과 임상 시험은 오랜 시간과 막대한 비용이 필요한 과정으로써 실패할 경우 큰 손실을 보게 된다. 약물 발견 과정에서는 이러한 손실을 최소화하기 위해 높은 효용성, 낮은 독성, 높은 효능을 가질 것으로 예상하는 화합물들을 신약 후보군으로 선택한다. 따라서 효과적인 신약 후보군의 선정은 이후 동물 실험과 임상 시험에 적용할 화합물의 수를 줄이게 되어 신약 개발 기간과 비용을 크게 단축할 수 있다.

Top-k 질의는 1 개 이상의 속성(attribute)들을 입력값으로 가지는 임의의 랭킹(ranking) 함수를 이용하여 함숫값이 가장 큰(혹은 작은) k 개의 튜플(tuple)들을 요청하는 질의이다. 임의의 x, y 에 대하여 x<= y 일 때, 항상 f(x) <= f(y)를 만족시키는 함수 f 를 단조증가함수라고 하며, 반대로 항상 f(x) >= f(y)를 만족할 경우 f 를 단조감소함수라고 한다. 어떤 함수 f 가 단조증가함수 또는 단조감소함수일 경우 f 를 단조함수라고 한다. 잘 알려진 임계값 알고리즘(threshold algorithm) [1, 2, 3]을 비롯하여 대부분의 top-k 질의 처리 기법들은 랭킹함수가 단조함수여야 하는 제약을 가진다.

약물 유사도는 두 화합물의 분자량, 수용성 및 지용성의 정도, 분자 면적 등의 특징이 얼마나 유사한지를 측정하는 지표로, 약물 유사도를 랭킹 함수로 사용하는 top-k 검색은 약물 발견 과정의 효율성을 높일 수 있다. 이미 약물로 쓰이는 화합물과의 약물 유사도를 활용해 특정 약물과 유사한 화합물을 찾아낼 수 있으며, 원하는 속성값을 갖는 가상의 화합물과의 유사도를 활용하여 원하는 특성을 갖는 화합물을 찾아낼 수도 있다. 그러나 유클리드 거리(Euclidean distance), 맨해튼 거리(Manhattan distance), 코사인 유사도(cosine similarity) 등 대부분의 유사도 함수는 비단조 함수이므로 전통적인 top-k 검색 시스템은 약물 유사도를 이용한 신약 발견에 사용할 수 없다.

본 논문에서는 본 연구실에서 개발 중인 비단조 랭킹함수를 지원하는 top-k 질의 처리 엔진을 활용하여 신약 개발 후보군 중 약물 유사도가 가장 높은 화합물들을 검색하는 엔진을 개발하였다.

본 논문의 구조는 다음과 같다. 2 절에서는 검색 엔진의 데이터셋으로 사용된 ChEMBL 데이터베이스[4]를 설명하며, 3 단원에서는 사용된 약물 유사도 함수를 설명한다. 4 절에서는 검색 엔진의 구조와 수행 방식을 소개하며, 5 절에서는 본 논문을 맺는다.

2. ChEMBL 데이터베이스

ChEMBL 데이터베이스[4]는 생물체에 작용하는 작은 화합물들의 정보들로 이루어진 데이터베이스이다. 이 데이터베이스는 화합물들의 2 차원 구조와 분자량, 지용성 지표, 결합 정보를 포함하여 화합물의 다양한 특징들로 구성되어 있다.

ChEMBL 데이터베이스(버전 22.1)에는 중복을 제외하고 총 1,686,695 개의 화합물들이 존재 하는데 이 중 질의 계산에 필요한 속성 값을 가지고 있지 않거

이 논문은 2017 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2014R1A2A2A01004454)

나, 잘못된 값을 가진 화합물들을 제외하면 총 1,641,322 개의 화합물이 있다. 이 1,641,322 개 중 약물에 사용되고 있는 화합물은 1,601 개가 존재하며 그 외 약물에 사용되지 않은 1,639,721 개의 화합물들이 약으로 사용된 1,601 개의 화합물 중 하나를 입력으로 주었을 때 유사한 화합물을 찾는 후보군으로 사용된다.

ChEMBL 데이터베이스에 존재하는 화합물들은 각자 ID 와 화합물의 다양한 특징과 관련된 23 가지의 속성을 가지고 있다. 하지만 유사도 측정에는 이들 중 다른 속성으로부터 계산되는 것들을 제외하고, 유사한 의미의 속성이 여러 개가 있는 경우 가장 대표적인 속성 하나만 선택하여 사용하였다. 이러한 조건으로 선택된 속성은 Molecular Weight(분자량), AlogP(지용성 지표), HBA(수소 결합 수용체의 수), HBD(수소 결합 공여체의 수), PSA(극성을 가진 원자들의 표면적의 합), 그리고 Heavy Atoms(수소를 제외한 원자들의 개수) 이렇게 7 가지이다.

3. 약물 유사도

본 논문에서는 유클리드 거리로 정의된 약물 유사도를 top-k 질의 처리 엔진의 랭킹함수로 사용하였다. 사용된 약물 유사도의 수식은 아래 (1)과 같고, top-k 검색 엔진은 함숫값이 가장 작은 k 개의 화합물을 찾는다. 여기서 c 와 q 는 각각 후보 화합물과 질의로 입력받은 약물에 사용되는 화합물이고, c_k, q_k 는 화합물 c 와 약물 q 의 k 번째 속성, 그리고 σ_k는 후보군 화합물들의 k 번째 속성의 표준편차를 의미한다.

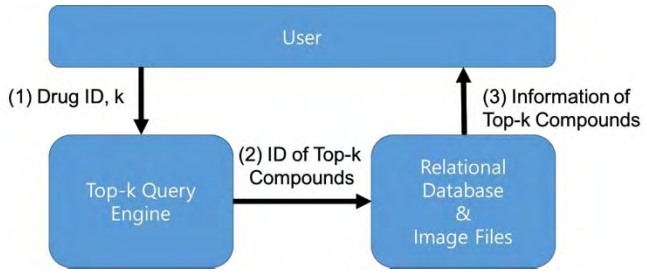
$$Dist(c,q) = \sqrt{\sum \left(\frac{c_k - q_k}{\sigma_k}\right)^2} \quad (1)$$

(1)은 각 속성의 중요성을 고려하여 가중치(weight)를 설정하지 않고, 표준편차를 이용해 모든 속성을 표준화(standardization)함으로써 각 속성이 유사도에 비슷한 기여를 하도록 한다. 적합한 가중치를 설정하거나 의미 있는 유사도 함수를 찾는 일은 본 연구의 주된 목적이 아니다. 하지만 본 연구실에서 개발한 top-k 질의 처리 엔진은 단조함수뿐만 아니라 비단조 함수도 지원하기 때문에 화합물의 속성을 입력값으로 사용하는 임의의 유사도 함수로 랭킹함수를 변경할 수 있으며 가중치의 변경도 자유롭다.

4. 검색 엔진 구조 및 수행 방법

본 논문에서 개발한 검색 엔진의 구조는 그림 1 과 같고, 실제 개발한 엔진에서 실행된 결과는 그림 2 와 같다. 그림 2 의 녹색 부분에 사용자가 유사도를 검색할 약물에 사용되는 화합물의 ID 와 k 값을 입력하면 top-k 질의 처리 엔진이 (1)의 랭킹함수 값이 가장 작은 k 개의 화합물들을 찾는다. 질의 결과에서 top-k 화합물들의 ID 를 이용하여 화합물의 ID 와 모든 속성을 저장하고 있는 관계형 데이터베이스(relational database)에 질의를 요청해 화합물들의 정보를 가져오며 ID 를 이름으로 갖는 화합물의 2 차원 구조 이미지 파일과 함께 사용자에게 보여준다. 그림 2 에서 빨간색과 파란색으로 표시된 부분이 각각 사용자가 입력한 화합물과 검색된 화합물들의 정보이다. 검색된 결과는 순차 탐색(sequential search) 방법으로 찾은 결과와 비교하여 정확성(correctness)를 검증하였다.

총 1,641,322 개의 후보 화합물을 사용했음에도 질의 시간이 상당히 짧다. k 가 5, 10, 20 일 때 질의에 각각 약 1.46 초, 1.53 초, 1.65 초가 소요되었다. 따라서 이 검색 엔진을 실제 신약 개발에 활용한다면 검색 속도에 의한 불편함 없이 사용할 수 있을 것이다.



(그림 1) 검색 엔진의 구조

(그림 2) 개발한 검색 엔진의 검색 결과 화면

5. 결론

본 논문에서는 본 연구실에서 개발한 비단조 랭킹함수를 지원하는 top-k 질의 처리 엔진을 활용하여 신약 발견을 위한 top-k 검색 엔진을 개발하였다. ChEMBL 데이터베이스에서 신약 후보군이 되는 화합물의 정보를 데이터셋으로 활용하였으며, 각 속성을 표준화한 유클리드 거리로 정의된 약물 유사도를 랭킹함수로 사용하였다. 개발된 엔진은 임의의 약물 유사도를 사용하도록 변경할 수 있으며 질의 시간도 매우 짧으므로 실제 신약 개발 단계에서 유용하게 활용될 수 있을 것이다.

참고문헌

[1] R. Fagin. Combining fuzzy information: an overview. SIGMOD Record, 31(2):109-118, 2002
[2] R. Fagin. Fuzzy queries in multimedia database systems. In PODS, pages 1-10, 1998.
[3] R. Fagin A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. In PODS, 2001.
[4] https://www.ebi.ac.uk/chembl/