

04

신약 개발 후보물질 발굴을 위한 거대 화합물 라이브러리 탐색 최적화

소속 정보컴퓨터공학부

분과 A

팀명 이성희

참여학생 성가빈, 이정민, 최우영

지도교수 송길태

과제 개요

과제 선정 배경

신약 개발에서 수억 개의 화합물 중 타겟 단백질과 결합 가능한 후보물질을 발굴하는 과정은 비용과 시간이 지나치게 많이 소요되어 효율성이 낮다.

과제 목표

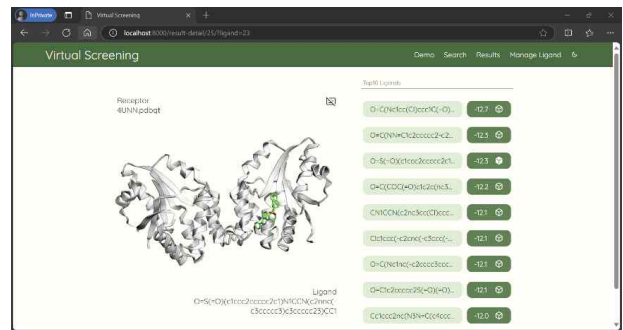
거대 화합물 라이브러리에서 주어진 타겟 단백질과 결합 확률이 높을 것으로 예측되는 화합물을 효율적으로 탐색하고 선별한다.

과제 내용

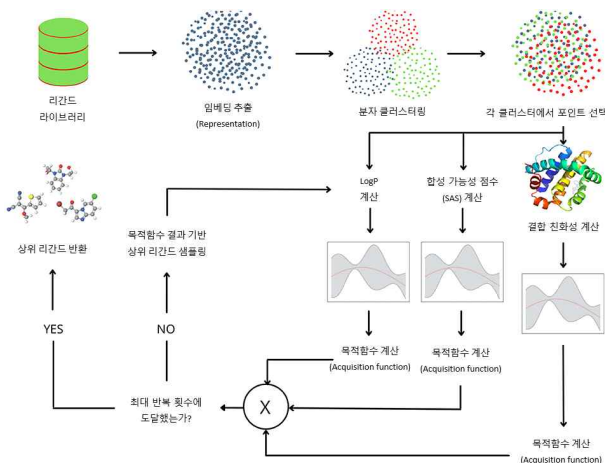
Clustering

k-means clustering

1. k개의 초기 중심점을 생성
2. 가장 가까운 중심점을 기준으로 클러스터에 배치
3. 클러스터 중심점 다시 계산
4. 중심점의 위치가 변하지 않을 때까지 반복



MEMES



1. 표현 방법에 따라 임베딩 추출
2. 초기 다양성을 위한 클러스터링 진행 후 초기 샘플 선택
3. 목적함수에 따른 가우시안 프로세스 회귀 정의
4. 베이지안 최적화 반복 수행
5. 선별된 리간드 중 상위 리간드 반환

분자 표현 방식

Fingerprint

분자 구조의 특정 부분구조나 성질의 유무를 비트 벡터로 나타내 분자 구조를 벡터로 표현

mol2vec

분자 구조를 단어처럼 취급하여 Word2Vec 알고리즘을 적용함으로써 분자 구조를 벡터로 표현

Acquisition function

EI (Expected Improvement)

개선된 값을 기대하는 정도 계산

PI (Probability of Improvement)

현재 최적값을 넘어설 확률 계산

UCB (Upper Confidence Bound)

상한 신뢰 구간 계산

Greedy

예측된 평균값만 사용해 그리디 방식

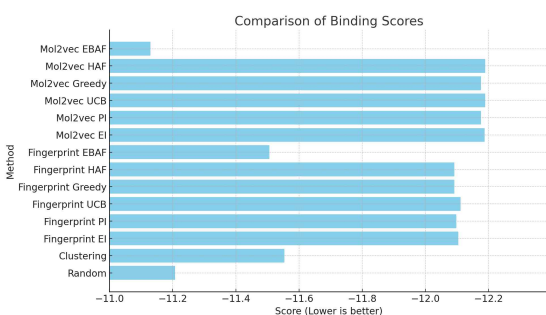
HAF (Hybrid)

EI와 UCB를 조합한 하이브리드

EBAF (Entropy-Based)

엔트로피 기반 샘플링 전략 사용

과제 결과



- 전체 200만개의 리간드 중 1%만 결합해보는 것으로 결합도가 높은 리간드를 찾아낼 수 있었다.
- 가장 좋은 결과를 낸 것은 MEMES에서 mol2vec로 분자를 표현하고 acquisition function으로 UCB를 사용했을 때이다.
- 신약 개발 과정에서 후보 물질 탐색 시간과 비용을 크게 절감해 약 개발의 효율성과 성공 가능성을 높이기를 기대할 수 있다.