신약 개발 후보물질 발굴을 위한 거대 화합물 라이브러리 탐색 최적화

착수보고서

08 - 이성최

202155565 성가빈

201924656 이정민

201845928 최우영

목차

1. 과제 배경 및 목표

1.1 과제 배경

신약 개발은 막대한 비용과 오랜 기간이 소요되는 어려운 과정이다. 그중에서도 가장 중요하면서도 큰 난관으로 작용하는 단계는 효과적인 후보물질을 발굴하는 것이다. 수억 개가 넘는 거대한 화합물 라이브러리에서 특정 타겟 단백질과 잘 결합할 수 있는 화합물을 찾아내는 일은 실로 시간이 많이 소요되는 과정이다.

전통적으로는 모든 화합물과 타겟 단백질 간의 결합 여부를 실험실에서 하나씩 실험을 통해 검사하는 방식을 사용해왔다. 하지만 이러한 접근 방법은 비용과 시간이 지나치게 많이 들어 효율성이 현저히 낮은 문제가 있다. 수억 개의 화합물을 모두 실험해보기에는 현실적인 한계가 있었다.

이런 상황에서 컴퓨터 기반의 가상 스크리닝 기술과 인공지능 기법이 등장하면서 상당한 진전이 있었다. 가상 스크리닝 기술은 화합물과 타겟 단백질 간의 상호작용을 컴퓨터 시뮬레이션을 통해 예측하는 기술로, 기존 실험실 기반 방식에 비해 시간과 비용 측면에서 획기적인 절감 효과가 있었다. 그러나 이러한 절감에도 불구하고 거대한 규모의 화합물 라이브러리를 매번 전부 탐색하기에는 여전히 오랜 시간이 소요되는 문제가 있다.

1.2 과제 목표

본 과제의 목표는 거대 화합물 라이브러리에서 주어진 타겟 단백질과 결합 확률이 높을 것으로 예측되는 화합물을 효율적으로 탐색하고 선별하는 것이다. 이를 위해 다양한 인공지능 및 최적화 알고리즘 방법을 제시하고 구현하여, 최대한 빠른 탐색 속도, 높은 정확성, 낮은 유사도를 얻는 것이 목표이다. 따라서 기존의 결합 확률 예측 모델에 힐 클라이밍, 시뮬레이티드 어닐링 기법 등 휴리스틱 탐색 기법을 개량하고 적용하여 탐색 과정을 가속화하고자 한다.

2. 대상 문제 및 요구조건 분석서

2.1 문제 분석

- **거대한 화합물 라이브러리의 크기**
  화합물 라이브러리의 규모가 수억 개에 달하기 때문에, 이를 전부 탐색하는 것은 현실적으로 불가능하다. 따라서 효과적으로 탐색 공간을 축소하고 유망한 화합물을 빠르게 찾아내는 전략이 필요하다.

- **탐색 결과의 다양성 확보**
  탐색 과정에서 발견된 화합물들이 서로 유사한 구조를 가질 경우, 실제 실험 단계에서 효과가 없을 가능성이 높아진다. 따라서 구조적 다양성을 유지하면서도 결합 확률이 높은 화합물을 선별하는 것이 중요하다.

- **알고리즘의 확장성과 효율성**
  사용되는 인공지능 및 최적화 알고리즘은 거대한 화합물 라이브러리에 대해서도 확장 가능하고 효율적으로 동작해야 한다. 따라서 알고리즘의 시간 및 공간 복잡도를 고려하여 설계하고 구현해야 한다.

- **도메인 지식**
  화합물과 타겟 단백질의 상호작용을 예측하고 해석하는 데에는 생물학, 화학 등 관련 도메인의 전문 지식이 필요하다. 따라서 도메인 지식을 학습하여 결과를 해석하는 과정이 필요하다.

2.2 요구조건 분석

2.2.1. 주요 요구조건

- **반응성이 높은 물질 선별 및 표시**
  추천 시스템은 타겟 물질에 대해 반응성이 높은 화합물들을 효과적으로 예측해 결과를 표시할 수 있어야 한다.

- **프로세스의 속도**
  스크리닝 과정은 10억여개의 데이터를 빠르게 처리할 수 있어야 한다.

- **화합물의 다양성 보장**

  추천 시스템은 비슷한 화학적 성질을 가진 화합물들만을 제안하는 것이 아니라, 다양한 화학적 성질을 가진 화합물을 포괄적으로 추천할 수 있어야 한다. 이는 신약 개발 과정에서 예상치 못한 부작용으로 인해 대체 물질이 필요할 경우에 대비하기 위해서이다.

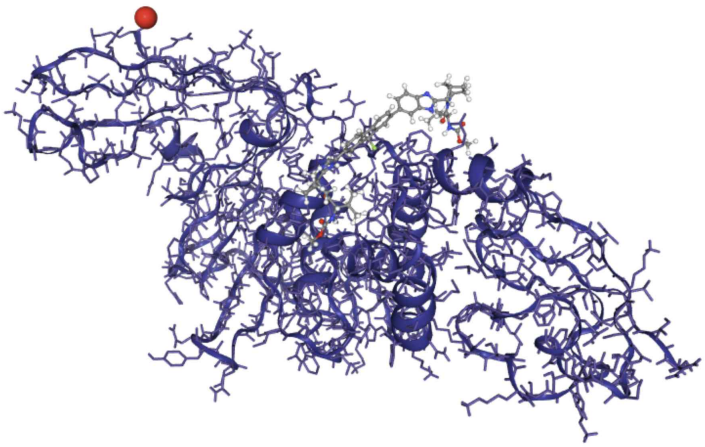
2.2.2. 추가적 요구조건

- **사용자 친화적 인터페이스**

  연구원이 쉽게 접근하고 사용할 수 있는 직관적인 사용자 인터페이스를 제공하여야 한다.

- **물질 결합 시각화**

  타겟 물질과 화합물이 결합되는 모습을 직관적으로 확인하기 위해 그래픽을 제공한다.



[그림] 단백질과 화합물의 결합 모습

3. 제약 사항 분석 결과 및 대책

3.1 제약 사항

- **처리 시간과 연산비**
  10억 개의 화합물을 모두 스크리닝하기 위한 시간과 비용은 매우 크다.

3.2 대책

- **병렬 처리와 분산 처리 기술 활용**
  화합물들을 여러 부분으로 나누어 동시에 처리함으로써 전체 작업의 완료 시간을 단축시킨다.

- **랜덤 테스트 추천**
  랜덤하게 화합물을 선택한 후 결합도 예측 모델에 넣은 후 score가 높으면 추천한다.

- **화합물 클러스터링 및 계층적 스크리닝**
  화합물을 미리 화학적 성질 등을 기준으로 비슷한 물질끼리 그룹과 대표 화합물을 설정한 후, 높은 score를 보이는 대표 화합물의 그룹은 추가로 스크리닝하고 나머지는 탈락시킴으로써 효율을 높인다.

4. 설계 문서

4.1 연구 진행 방향

1. **개발 환경 구성**
   개발 환경을 구성한다. 10억개의 화합물 데이터가 있으나, 우선은 그 중 10만개 정도만 사용한다. 표적 단백질과 화합물 사이의 결합에 대한 점수를 측정하는 모델을 사용할 것이다.

2. **랜덤으로 화합물 추천**
   랜덤으로 100개 정도의 화합물을 선택해서, 그 중 점수가 높은 10개의 화합물을 추천해 준다.

3. **clustering하여 화합물 추천**
   화합물들을 clustering하고 시작한다. 각 cluster마다 대표할 화합물을 몇 개씩 선정하고, 점수를 측정한다. 평균적인 점수가 낮은 cluster는 제외한다. 점수가 높은 cluster 중에서 다양한 cluster의 화합물을 추천한다.

4. **알고리즘 분석 & 모델 개선**
   위 방법들의 문제점을 분석하고, 화합물을 더 빠르고 효율적으로 탐색할 수 있는 적합한 알고리즘을 생각한다. 추가적으로, 목표를 달성하기 위하여 모델을 개선한다.

5. **시각화**
   웹 상에서 표적 단백질에 따른 추천되는 화합물을 확인할 수 있도록 한다. 그리고 단백질과 화합물의 결합을 시뮬레이션한다.

4.2 개발 환경 및 사용 기술

4.2.1. 개발 환경

- python
- 화합물 데이터 라이브러리
- 스코어 측정 모델
- JavaScript

4.2.2. 사용 기술

- **bayesian optimization**

  베이지안 최적화(bayesian optimization)는 알려지지 않은 목적 함수의 값을 최대 또는 최소로 만드는 입력 값을 찾는 것을 목표로 한다. 이 방법은 하이퍼파라미터 최적화에 주로 사용되며, 이 경우 하이퍼파라미터가 입력 값이 된다.

  베이지안 최적화는 surrogate model과 acquisition function으로 구성되고, 이 두 과정을 반복하면서 목적 함수가 최대 또는 최소가 되는 지점을 찾게 된다.

  Surrogate model은 이전의 입력 값과 그에 따른 목적 함수의 값들을 사용해 목적 함수를 확률적으로 추정한다. Gaussian process가 surrogate model로 사용될 수 있다.

  Acquisition function은 surrogate model의 확률적 추정을 바탕으로 다음에 사용할 입력 값들을 추천한다. 주로 expected improvement가 사용된다.

- **active learning**

  active learning은 모든 데이터를 라벨링해서 학습하는 대신, 분류가 완료된 일부 데이터로 시작해 모델이 라벨링이 필요할 때 요구하도록 한다. 데이터들을 분류하는 데에는 많은 비용이 소모되는데, active learning을 이용하면 이를 줄일 수 있다.

  먼저 일부의 라벨링 된 데이터를 가지고 모델을 학습시킨다. 학습된 모델은 분류되지 않은 데이터들의 클래스를 추측하고, 그 중 가장 필요한 데이터에 대해 라벨링을 요청한다. 그렇게 추가된 데이터로 위의 과정을 반복하게 된다.

  어떤 데이터가 필요한 데이터인지 판단하는 방법에는 여러 방법이 있다. uncertainty sampling 방식은 추측한 결과가 가장 불확실한 데이터부터 라벨링을 요청한다. 그리고 query by committe 방식에서는 여러 개의 모델을 사용하는데, 각 모델이 추측한 결과가 가장 일치하지 않는 데이터부터 라벨링을 요청한다. 그 외에도 co-training 기법을 활용하거나 learning loss를 기준으로 요청할 데이터를 선정할 수 있다.

5. 일정 및 역할 분담

5.1 추진 체계 및 일정

5월		6월				7월					8월					9월				10월		
4	5	1	2	3	4	1	2	3	4	5	1	2	3	4	5	1	2	3	4	1	2	3
착수 보고서 작성																						
		개발 환경 구축																				
			완전 탐색 및 무작위 탐색 구현																			
				clustering 구현																		
						알고리즘 분석 및 개선																
						웹 UI 구축																
										중간 보고서 작성												
														알고리즘 분석 및 개선								
																시각화						
																			최종 보고서 작성			

5.2 구성원 역할 분담

이름	역할 분담
성가빈	개발 환경 구축 및 관리 시각화
이정민	개발 환경 구축 및 관리 웹 서버 개발
최우영	시각화 웹 서버 개발
공통	알고리즘 구현 알고리즘 평가 보고서 작성

6. 참고 자료 및 출처

- [그림 1] DockThor (lncc.br)

  해당 이미지의 저작권은 LNCC에 있습니다.

- 베이지안 최적화(Bayesian Optimization) (tistory.com)

- [ML] 베이지안 최적화 (Bayesian Optimization) (tistory.com)

- 김호찬 and 강민제. (2020). Comparison of Hyper-Parameter Optimization Methods for Deep Neural Networks. 전기전자학회논문지, 24(4), 969-974.

- 송하윤, 남세현. (2020). 베이지안 최적화를 이용한 이동 경로 예측 모델의 성능 개선. 한국정보처리학회 학술대회논문집, 27(2), 846-849.

- ML | Active Learning - GeeksforGeeks

- [Seminar] Active Learning – DSBA (korea.ac.kr)

- 강재호, 류광렬, 권혁철. (2005). 능동적 학습을 위한 군집화 기반의 다양한 복수 문의 예제 선정 방법. 지능정보연구, 11(1), 169-189.