단백질-리간드 결합 부위 및 친화도 예측 모델 연구

팀명:
으쌰으쌰

부산대학교
정보컴퓨터공학
부

202255512 김다현
202255552 박주은
202255565 안수현

지도교수: 송길태

목차

1. 과제 배경 및 목표

a. 과제 배경

b. 과제 목표

2. 요구사항 분석

a. 기능적 요구 사항 분석

b. 비기능적 요구 사항 분석

3. 개발 환경 및 사용 기술

a. 개발 환경

b. 사용 기술

4. 현실적 제약 사항

5. 개발 일정 및 역할 분담

a. 개발 일정

b. 역할 분담

6. 참고문헌

1. 과제 배경 및 목표

a. 과제 배경

단백질과 리간드 간의 결합 부위(binding site)와 결합 친화도(binding affinity)를 예측하는 것은 신약 개발에서 중요한 과제이다. 예를 들어, 약물 분자가 특정 단백질 수용체에 결합해 효과를 발휘하거나, 호르몬이 세포 표면의 수용체에 신호를 전달하는 과정 모두가 단백질-리간드 상호작용에 기반한다. 이러한 결합 특성을 정확히 예측하면 실험적 검증에 드는 비용과 시간을 줄이고, 후보 물질의 탐색 효율을 크게 높일 수 있다.

또한 결합이 발생하더라도, 그 결합이 약하고 불안정하다면 약물로서 효과를 기대하기 어렵다. 따라서 얼마나 강하고 안정적으로 결합이 형성되는지 알 수 있는 결합 친화도를 예측하는 것도 중요하다. 이 두 정보를 함께 예측함으로써, 가능성이 높은 약물 후보를 보다 정확하게 선별할 수 있다.

기존에는 단백질과 리간드의 결합 부위를 예측할 때 서열(sequence) 기반 접근법 보단 3차원(3D) 구조 기반 접근법이 더 정확한 예측 성능을 보여왔다. 하지만 이런 접근법을 활용한 모델은 benchmark 데이터에선 높은 성능을 보였으나 그 외에 현실에서 적용할 때는 그렇지 않았다. 학습과 추론에 필요한 3D 구조 데이터가 부족하여 일반화에 낮은 성능을 보인 것이다. 이 때문에 구조에 의존하지 않는 기술의 서열(sequence) 기반 예측 기법의 필요성이 점점 커지고 있다.

최근에는 딥러닝 기술의 발전으로, 단백질 아미노산 서열 정보만으로도 의미 있는 결합 정보를 추출하는 것이 가능해졌다. 특히, ProtTrans, ESM 등의 모델을 활용한 단백질 서열 임베딩 기법과, 화합물의 SMILES 문자열 또는 그래프 표현 기반 임베딩 발전으로 인해, 3D 구조 없이도 생화학적 의미를 포착할 수 있는 기반이 마련되고 있다.

이에 따라, 단백질 서열과 리간드 구조 정보를 입력으로 하여, 결합 부위 그리고 결합 친화도를 예측하는 서열 기반 딥러닝 모델에 대한 연구가 활발히 진행되고 있다.

b. 과제 목표

본 과제에서는 단백질과 리간드 간의 상호작용을 정밀하게 이해하고자, 단백질과 리간드 간의 결합 부위(binding site)와 결합 친화도(binding affinity)를 예측할 수 있는 딥러닝 기반의 예측 모델을 개발하는 것을 목표로 한다.

또한 이렇게 개발된 모델을 쉽게 활용할 수 있도록 웹페이지 형태의 인터페이스로 구현하여, 사용자들의 접근성과 편의성을 향상시키는 것을 목표로 한다. 세부 목표는 다음과 같다.

1. 단백질 서열 기반 구조 및 기능 정보 학습
2. 결합 부위 및 결합 친화도 예측 모델 구현

3. 모델 활용을 위한 웹 기반 인터페이스 구현

2. 요구사항 분석

a. 기능적 요구 사항 분석

1. 사용자 로그인
    - 사용자가 웹 페이지 인터페이스에 로그인할 수 있어야 한다.

2. 사용자 입력
    - 사용자가 단백질 서열(sequence)과 리간드 구조를 웹 페이지 인터페이스 상에서 입력할 수 있어야 한다.

3. 결과값 처리
    - 사용자가 입력한 단백질과 리간드 사이의 결합 부위(binding site)를 결과값으로 도출해야 한다.
    - 사용자가 입력한 단백질과 리간드 사이의 결합 친화도(binding affinity)를 결과값으로 도출해야 한다.

4. 결과값 시각화
    - 웹 페이지 인터페이스를 통해 단백질과 리간드 사이의 결합 부위를 시각적으로 보이게 해야한다.

5. 결과값 저장

- 단백질과 리간드 사이의 결합 부위 및 결합 친화도 값을 저장할 수 있어야 한다.

6. 저장된 값 보기
    - 저장해두었던 단백질과 리간드 사이의 결합 부위 및 결합 친화도 값을 한 번에 볼 수 있어야 한다.

b. 비기능적 요구 사항 분석

1. 성능
    - 모델 예측 속도 : 사용자의 입력에 대해 30초 이내에 결과값을 도출해내야 한다.
    - 시스템 응답 시간 : 모델 예측 후, 웹 페이지 인터페이스에 1초 이내에 확인 가능해야한다.

2. 보안성
    - 로그인 : 로그인되지 않은 사용자는 예측 모델을 사용할 수 없다.
    - 정보 보호 : 다른 사람이 예측을 시도하거나, 저장한 단백질-리간드 쌍에 대한 정보를 타인은 절대 볼 수 없다.

3. 안정성
    - 예외 처리 : 비정상적인 입력(ex. 너무 긴 서열, 비표준 아미노산)에 대해 에러 메시지를 반환하고, 예측을 하지 않는다.

- 자원 제어 : 예측 요청이 과도하게 쌓일 경우, 서버 과부하를 막기 위해 예측을 제한한다.

4. 가용성
    - 접근성: 사용자는 다른 프로그램의 설치 없이도, 웹 브라우저를 통해 웹 페이지 인터페이스에 접근 가능해야 한다.
    - 서비스 시간: 시스템은 24시간동안 운영 되어야 하며, 점검 시에는 이를 미리 공지 해야한다.

3. 개발 환경 및 사용 기술

a. 개발 환경

1. 프레임워크: PyTorch
2. 데이터셋:
    - 학습용 - PDBbind
    - 테스트용 - COACH420
3. 전처리 도구: RDKit
4. 시각화 도구: TensorBoard

b. 사용 기술

1. 모델 개발
    - CNN + Attention
2. 웹 페이지 개발
    - 프론트엔드 : React

- 백엔드 서버 : FastAPI
- API 통신 : RESTFul API
- 시각화 : NGL Viewer

4. 현실적 제약 사항

1. Binding pose 정보 반영 부재
    - 모델의 입력 및 처리 과정이 시퀀스 기반으로 구성되어 있어 실제 결합 자세(binding pose)나 구조적 적합성을 직접적으로 반영하기 어렵다. 이는 실제 결합에서 중요한 구조적 적합성 및 상호작용을 충분히 고려하지 못하게 하여 예측 정확도 및 구조 기반 해석력 측면에서 한계로 작용할 수 있다.

2. 물리화학적 요인 반영 어려움
    - 수소결합, 소수성 상호작용, 원자 간 거리 및 전하 분포 등 실제 결합에 영향을 주는 세밀한 물리화학적 요인을 모델에 반영하여 실제 결합 환경에서의 상호작용을 정밀하게 예측하는 데 어려움이 있다.

3. Induced Fit 효과 미반영
    - 실제 단백질-리간드 결합 과정에서는 리간드의 결합에 따라 단백질 구조가 유연하게 변형되는 Induced Fit 현상이 발생할 수 있다. 그러나 본 모델은 고정된 서열 정보를 기반으로 예측을 수행하기 때문에 결합 시 발생하는 구조적 재배열이나 유연성

변화를 반영하지 못하여 실제 결합 상황 및 결합 친화도 예측에 차이를 불러올 수 있다.

4. pH, 이온 농도 등 생체 환경 요소 고려 어려움
   - 실제 생체 내 단백질-리간드 결합은 pH, 이온 농도, 온도, 수용성 환경 등 다양한 생리학적 조건의 영향을 받지만 본 모델을 이러한 환경적 요소를 입력으로 포함하지 않기 때문에, 실제 조건 하에서의 결합 특성 변화를 반영하기 어렵다.

5. 개발 일정 및 역할 분담

a. 개발 일정

														최종 보고서

b. 역할 분담

공통	- 데이터 전처리 및 임베딩 - 모델 구조 구현 및 학습 - 모델 성능 평가 - API 설계 및 프론트엔드 연동
김다현	- 웹 인터페이스 UI 구성
박주은	- 웹 인터페이스 API 연동
안수현	- 웹 인터페이스 시각화

6. 참고문헌

1. **Lee, I., & Nam, H. (2022).** Sequence-based prediction of protein binding regions and drug–target interactions. *Journal of Cheminformatics, 14*(5), Article 5. https://doi.org/10.1186/s13321-022-00584-w
2. **Seo, S., Choi, J., Choi, S., Lee, J., Park, C., & Park, S. (2024).** Pseq2Sites: Enhancing protein sequence-based ligand binding-site prediction accuracy via the deep convolutional network and attention mechanism. *Engineering Applications of Artificial Intelligence, 127*, 107257. https://doi.org/10.1016/j.engappai.2023.107257

3. **Li, K., Xiao, X., Zhong, Z., & Yang, G. (2025).** *Accurate and generalizable protein–ligand binding affinity prediction with geometric deep learning*. https://arxiv.org/abs/2504.16261