DP(Disease Prediction)

1. 프로젝트 개요

1-1. 프로젝트 목적

단백질 간 관계, 유전자 정보, 단백질 구조 등 다양한 생물학 데이터를 통합 분석하여 질병을 예측할 수 있다. 즉, 단백질의 생물학적 정보를 기반으로, 입력받은 단백질이 어떤 질병과 연관되어 있는지 예측하는 시스템이다.

1-2. 네트워크 연결망을 사용하지 않고, 왜 GCN을 사용하는가?

기존의 질병 예측 시스템에서 사용되는 전통적인 네트워크 연결망 기반 분석 방법은 주로 단백질 간의 연결(PPI), 유전자-질병 매핑 등을 기반으로 **정적 연결망(graph)** 을 구성 같은 방식으로 구성된다.

그래프 상에서 연결 강도(combined score 등)를 기준으로 이웃 탐색 또는 전파 기반 예측 수행할 수 있다. 주로 rule-based, random walk, 또는 통계적 네트워크 분석 방법 사용될 정도로 전통적이다. 그러나 이러한 전통적인 연결망 방식은 다음과 같은 한계를 가진다:

- **이웃 정보에만 의존하는 한계**: 네트워크 상에서 직접적으로 연결된 이웃의 정보에만 의존하여 예측하기 때문에, **간접적이고 복잡한 관계(2-hop 이상)** 는 잘 반영되지 않는다.
- **피처 정보의 결합이 어려움**: 단백질의 서열, 구조, 기능(GO term)과 같은 복잡한 속성 데이터를 네트워크 구조와 함께 통합적으로 학습하기 어렵다.
- **노이즈에 취약하고 일반화가 어려움**: 연결망에서의 한두 개의 잘못된 엣지나 결손된 정보가 전체 예측 성능에 영향을 줄 수 있으며, 데이터가 희소할수록 정확도 저하가 심각하다.

GCN(Graph Convolutional Network)은 그래프 구조를 학습에 직접 활용하는 딥러닝 모델로, 다음과 같은 방식으로 위의 단점을 극복한다:

전통 연결망의 한계	GCN의 해결 방식
직접 연결에만 의존	GCN은 k-layer를 통해 다중 hop 이웃의 정보를 자연스럽게 통합하여 간접적 관계도 반영
속성(feature) 통합 어려움	단백질 서열, 기능 정보 등 노드 feature를 함께 학습하며 관계와 특성의 통합적 표현 가능
노이즈 및 데이터 결손에 취약	학습된 노드 임베딩은 구조와 속성 모두를 반영하므로, 일부 연결이 누락되어도 강건한 예측 가능
지식 기반 의존	GCN은 데이터 기반 자동 학습 모델로, 사전에 설계된 룰 없이도 관계를 스스로 학습함

따라서 본 프로젝트는 단순한 네트워크 탐색 기반 접근이 아닌, GCN을 활용한 학습 기반 예측 방식을 채택함으로써 다음과 같은 장점을 확보한다:

- 단백질 간 복잡한 상호작용 및 질병 연관성의 깊이 있는 이해
- 다중 데이터(구조, 기능, 연관성 등)를 통합한 효율적인 표현 학습
- 기존 방식 대비 일반화 가능성과 확장성이 높은 질병 예측 시스템 설계 가능

GCN은 단순한 연결 유무가 아닌, **"누가 누구와 어떻게 연결되어 있으며, 어떤 생물학적 특성을 가지는가?"** 를 동시에 학습하여, 더 정밀하고 신뢰도 높은 질병 예측을 가능하게 한다.

---

2. 프로젝트 구조 요약

2-1. 사용자 INPUT

본 시스템은 단백질 기반 질병 예측을 목표로 하며, 사용자는 예측하고자 하는 단백질 정보를 입력으로 제공하게 된다. 입력 정보는 다음과 같은 형태로 제공된다:

항목	설명
`Protein ID or Name`	Protein ID (예: `9606.ENSP00000346839`), UniProt ID, 또는 단백질 이름
`Input Type` *(선택)*	미지의 단백질 또는 기존에 알려진 단백질 여부 (선택적)
`Feature Option` *(선택)*	기능, 구조, GO term 등 포함 여부 (향후 확장 가능)

> 시스템은 입력된 단백질 ID 또는 이름을 내부 단백질 식별자(Protein ID 또는 UniProt ID)로 변환한 뒤, 해당 단백질의 임베딩 벡터를 생성하고, 학습된 모델을 통해 관련 질병 예측 결과를 도출한다.

2-2. 모델 OUTPUT

모델은 입력된 단백질에 대해, 관련 질병 및 관련 단백질을 예측하고, 기능적 정보를 함께 제공한다. 예측 결과는 다음과 같은 항목으로 구성된다:

항목	설명
`related_diseases`	해당 단백질과 연관될 가능성이 높은 질병 리스트 및 score
`related_proteins`	PPI 기반으로 유사성이 높은 단백질 리스트 및 유사도 점수
`functional_annotation`	단백질의 기능 요약 텍스트 (GO term, 구조 정보 포함)
`network_graph`	질병 및 단백질을 포함한 시각화용 그래프 데이터 (nodes, edges)

> 모델은 위 정보를 바탕으로 질병 예측 결과를 사용자에게 텍스트 및 시각화 형태로 제공할 수 있도록 구성된다.

---

3. GCN 모델 학습 계획

3-1. GCN 학습에 사용될 데이터 Feature(수정 될 수 있음)

💬 기본 형태

> 아래는 질병 예측을 위한 GCN 학습에 사용되는 주요 데이터 컬럼이며, 단백질-질병-유전자 간의 관계 및 단백질 기능 정보를 통합하여 feature로 활용된다.

컬럼명	설명	기능 / 용도
`Disease ID`	질병 고유 ID (예: DOID_0050156)	분석 기준이 되는 질병 식별자

컬럼명	설명	기능 / 용도
`Disease Name`	질병 이름 (예: breast cancer)	결과 해석 및 시각화 시 사용
`Gene ID`	유전자 식별 번호 (NCBI 기준)	병합 키로 사용됨
`UniProt_ID`	단백질 고유 ID (UniProt 기준)	단백질 식별자
`GO_Terms`	단백질의 기능 분류 정보 (Gene Ontology)	기능 기반 분석에 활용
`PDB_IDs`	단백질 구조 ID (Protein Data Bank)	구조 기반 연구 시 / 질병 예측 활용
`PubMed_IDs`	관련 논문 ID	참고 문헌 추적 가능
`Protein_ID_Formatted`	`9606.` 접두어가 붙은 단백질 ID	PPI 병합 시 사용되는 표준 포맷
`protein1`, `protein2`	상호작용하는 단백질 쌍	PPI 네트워크 연결 정보
`combined_score`	단백질 간 상호작용 강도 (0~1000)	상호작용 신뢰도 기준

🧬 UniProt 메타데이터 컬럼 설명

> 아래 컬럼들은 UniProt 데이터베이스에서 제공하는 **기본 생물학적 정보**로, 단백질의 기능, 서열, 존재 근거 등을 포함한다. 본 프로젝트에서는 GCN 모델 학습을 위한 feature로 활용되며, 예측 결과의 생물학적 해석에도 사용된다.

컬럼명	설명	용도
`entryId`	UniProt 내부 entry ID	고유 식별자
`gene_x`	유전자 이름 (from UniProt)	분석에 사용되는 유전자명
`geneSynonyms`	유전자 별칭들	보조 검색 키
`isReferenceProteome`	기준 단백질 여부 (True/False)	표준 단백질 여부 확인
`isReviewed`	리뷰 여부 (Swiss-Prot vs TrEMBL)	데이터 신뢰도 판단
`sequenceChecksum`	시퀀스 무결성 검사용 코드	데이터 검증
`sequenceVersionDate`	시퀀스 버전이 갱신된 날짜	버전 추적
`uniprotAccession`	공식 단백질 Accession ID	고유 단백질을 추적 가능
`uniprotId`	단백질 이름 ID	uniprot DB에서 고유 식별 번호
`uniprotDescription`	단백질 기능 설명	분석에 사용 가능한 설명
`taxId`	생물종 ID (9606 = 인간)	종 분류
`organismScientificName`	생물학적 학명	Homo sapiens
`globalMetricValue`	글로벌 통계 수치	분석 또는 품질 관련 지표
`uniprotStart`, `uniprotEnd`	단백질 시퀀스의 일부분 위치	위치 기반 기능 연구 가능

컬럼명	설명	용도
uniprotSequence	단백질 아미노산 서열	생물학적 분석의 핵심
modelCreatedDate	이 데이터가 만들어진 날짜	데이터 생성 시점 추적
organismCommonNames	종의 일반 이름 (human)	종 구분 가능능
proteinFullNames	단백질 전체 이름	설명적 이름
latestVersion, allVersions	최신 및 전체 버전 기록	변경 이력 파악
isAMdata	Additional metadata 포함 여부	추가 분석용
organismScientificNameT	종 학명 (변형 또는 반복)	구조상 중복 필드 가능성
version	내부 DB 버전	시스템용
proteinShortNames	단백질 약칭	간단한 이름 사용 시
uniprotAccession_unchar	미지의 단백질 ID	기능 발견 전 데이터 가능성
entry_name	UniProt entry 이름	주로 유전자 기반 명칭
protein_name	단백질 이름 (일반적 표현)	단백질 이름름
organism	종 이름 (일반 표현)	Human
tax_id	종 ID (중복 필드 가능)	taxId와 유사
gene_y	유전자 이름 (병합된 또 다른 필드)	gene_x와 동일 가능
protein_existence	단백질 존재 근거 (1~5 등급)	신뢰도 판단
sequence_version	시퀀스 버전 정보	변경 이력 추적용
sequence	아미노산 서열 정보	단백질 기능 예측의 핵심