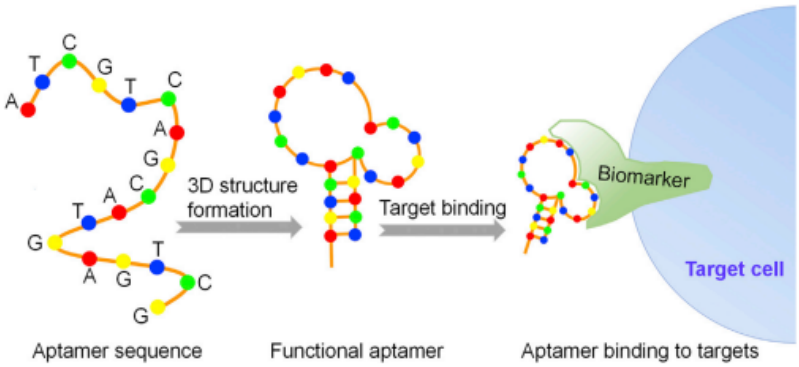
# 2022 전기 졸업과제 착수보고서

## 목차

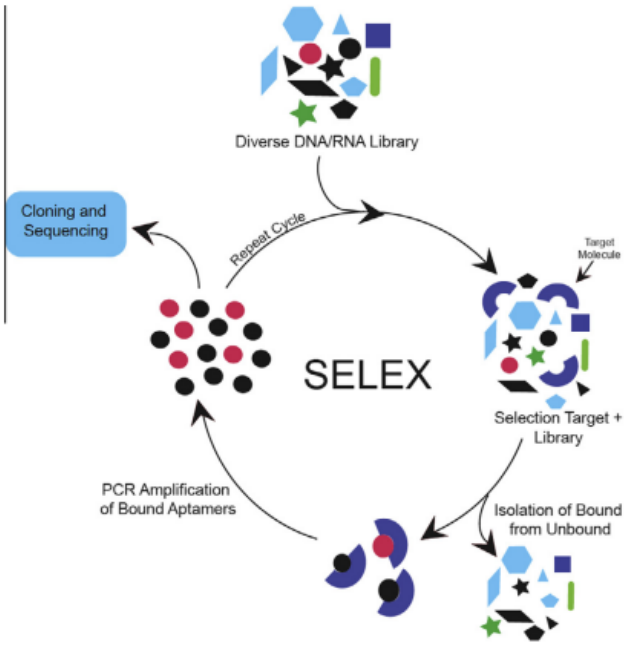
# 1. 과제 배경 및 목표

## 1. 과제 배경




[그림 1] Aptamer의 형태와 실제 생물학적 결합 과정[1]

Aptamer(이하 압타머)는 RNA/DNA와 유사한 구조를 가지는 핵산 물질로, 표적 단백질에 다양한 형태로 결합하기에 암 치료제 분야에서 사용되고 있다. 압타머는 항체와 비교될 수 있을 정도의 높은 결합 친화도를 가지기 때문에 항체의 대체 물질로서 각광받고 있다. [그림 1]과 같이 일련의 염기서열로 이루어진 압타머 시퀀스가 있으면, 이를 구조적으로 변형하여 온전한 압타머를 생성하고, 이후 표적 단백질과 결합하여 특정 성분을 전달하게 된다.




[그림 2] SELEX 기법의 전체 프로세스[2]

다양한 후보 압타머들 중에 가장 결합 친화도가 높은 것을 선택하기 위해서 제일 많이 사용되는 방법은 SELEX이다. SELEX는 유용한 유전 물질을 선택하는 과정을 [그림 2]와 같은 단계를 거쳐 반복 시뮬레이션을 통해 얻어내는 방식이다. 그러나 이 방법은 실제로 유전 물질을 합성 & 분리하기에 많은 노력과 시간을 필요로 한다. 따라서, 이를 단축시키기 위한 컴퓨터 시뮬레이션 시스템을 이용한 사례가 등장하였다. 가령, 다층 퍼셉트론(MLP)을 활용한 방식[3]이나, 합성곱

신경망(CNN)을 이용한 모델[4], 혹은 트리 기반 랜덤 포레스트(RF) 모델을 사용해서 예측을 수행하는 경우[5]가 있다.

본 과제에서는 그 중에서도 MLP, CNN 등의 인공신경망을 적용한 예측 모델을 생성하되, 기존에 개발된 모델과 성능을 비교하여 모델 최적화를 수행할 예정이다. 또한, 예측 과정에서 어떠한 변인이 주 영향을 끼치는지를 파악하고, 단백질에 결합할 가능성이 높은 압타머의 염기 서열 특징을 파악하여 일반화된 형태를 제시할 계획이다.

### 2. 과제 목표

본 과제에서는 인공신경망을 이용하여 표적 단백질에 대해 결합 친화도가 높은 압타머를 생성하는 예측 모델을 구현할 계획이다. 더 나아가서 압타머 생성 모델에서 어떤 변인이 영향력이 높은지를 파악 및 특정 단백질에 결합하는 압타머의 패턴을 효과적으로 도출하는 것을 목표로 한다.

- 인공신경망을 활용하여 주어진 단백질과 염기서열이 있을 때 결합 여부를 판별하는 모델 제작
    - MLP, CNN 등의 모델을 활용한 학습 모델 생성
- DeepBind, AptaNet 등의 기존 모델과 비교하여 성능 분석 및 모델 개선
    - 기존 모델을 토대로 구조 변형
    - 기존에 생성된 모델 데이터를 활용한 검증 및 성능 비교
- 결합 여부에 영향을 끼치는 변인과 결합 가능성 압타머의 패턴(Motif) 도출
    - 효과적인 결합 여부 판별을 위한 데이터 전처리 수행
    - 속성별 종속 변수에 관한 영향력을 비교 분석
    - 특정 단백질에 결합력이 높은 패턴을 분석 및 시각화

## 2. 문제 파악 및 요구사항 분석

### 1. 데이터 전처리

딥러닝 모델이 학습할 수 있도록 두가지 방법으로 데이터를 가공하고 재구성한다.

- 입력된 데이터의 A, G, T, C 네 가지 염기에 대한 정보를 딥러닝 모델이 학습 가능한 One-Hot Vector의 형태로 변환한다.
- 제시된 sequence를 염기 다양한 길이로 잘라 sequence를 생성한 후 One-Hot Vector의 형태로 변환한다.
- 모델 훈련과 성능 평가를 통해 가장 적절한 데이터 전처리 방법을 선택해야한다.
- 데이터의 불균형이 존재할 시 이를 해결할 데이터 샘플링 방법도 선택해야한다.

### 2. 적절한 하이퍼 파라미터

많은 훈련을 통해 적절한 하이퍼 파라미터를 찾아야한다.

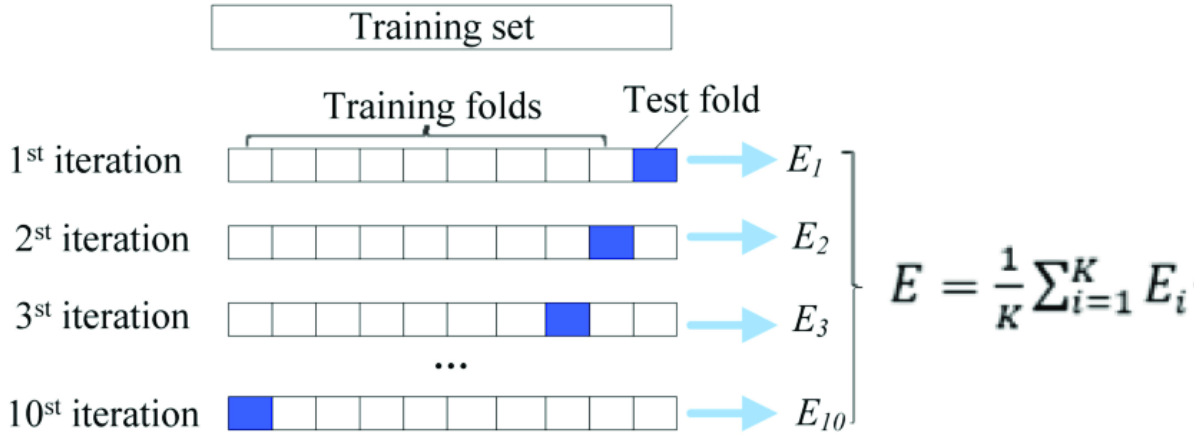
- 데이터 불균형이 있을 때 여러 데이터 샘플링의 하이퍼 파라미터들 또는 각 모델의 최적의 하이퍼 파라미터들을 찾아낸다.

- 과적합(overfitting)이 발생하지 않고 Traing Set 뿐만 아니라 Test Set에서도 최고의 성능을 낼 수 있는 학습 횟수(epoch)와 학습률(learning rate)를 찾아낸다.

## 3. 모델 검증

어떠한 데이터의 분포 속에서도 좋은 성과를 보여야하고 적절한 성과 지표를 사용해야한다.

- 불균형 데이터에서도 정확한 성능을 확인할 수 있는 AUC(Area Under the ROC)를 사용한다.
- [그림 3]과 같은 10-fold cross validation을 이용하여 Training set과 Validation set으로 분리하여 10번 테스트하여 성과를 확인해 어떤 데이터와 모델을 선택할지 정한다.




[그림 3] 10-fold cross validation의 과정

## 4. Motif 결합 가능성 예측 딥러닝 모델

Motif 판별자 딥러닝 모델은 학습을 통해 실제 데이터와 가짜 데이터를 구분할 수 있어야하고 Motif 생성자 딥러닝 모델은 표적 단백질에 대해 결합 가능성이 큰 Motif를 생성해야한다.

- CNN 모델을 기반으로 표적 단백질 데이터를 이용해 학습한다.
- 무작위로 Motif를 생성하는 알고리즘보다 효율적인 속도와 성능을 보여야한다.
- 최종적인 목표는 Deep Bind와 비슷하거나 그보다 뛰어난 판별 모델을 생성한다.

## 5. 모델 범용성

본 과제의 생성 모델은 특정한 표적단백질만이 아닌 다양한 표적단백질에 대해 유의미한 성능을 보여야한다.

- 다양한 표적 단백질의 데이터를 활용해 어떤 상황에서도 표적 단백질에 가능성이 큰 Motif를 생성한다.

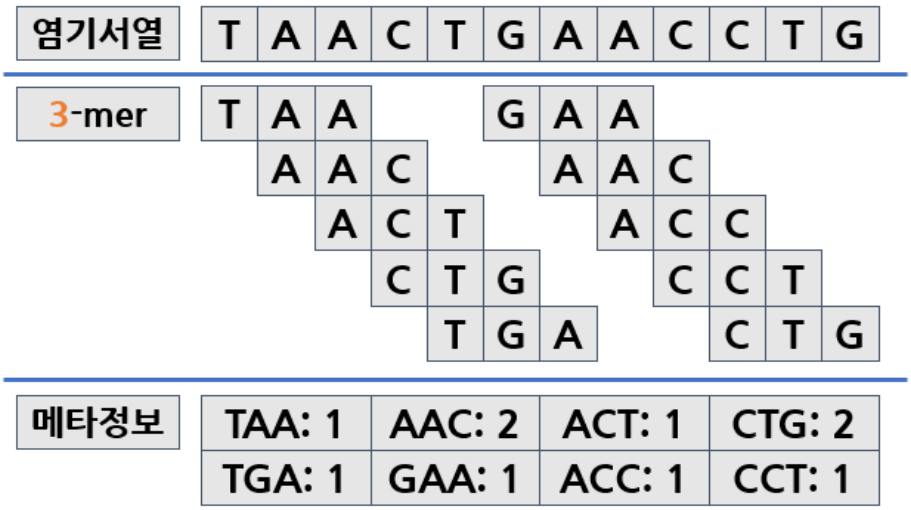
# 3. 상세 구현 방안

## 1. 개발 환경

- 개발 언어: Python3
- 활용 라이브러리: Scikit-learn, TensorFlow, Keras
- 활용 플랫폼: Jupter Notebook, Google Colab (안정적인 GPU 환경 제공)
- 주요 실행환경: Window, Mac

## 2. 주요 활용 기술

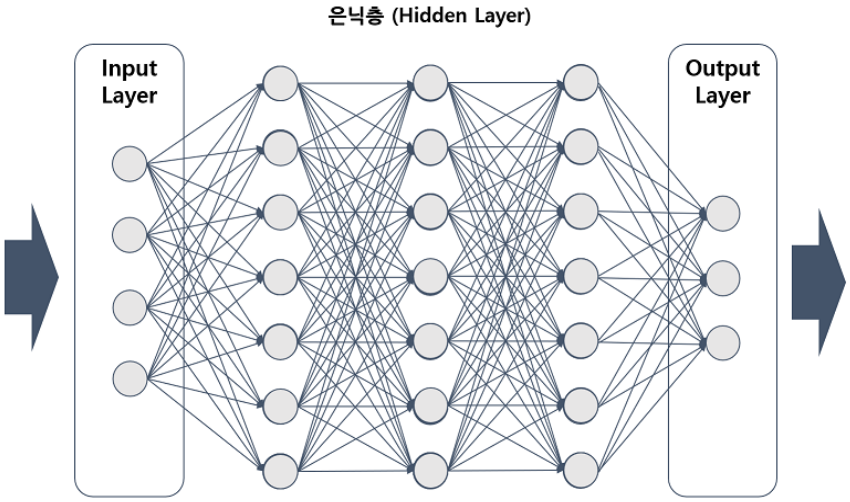
> k-mer




[그림 4] 주어진 염기서열에서 3-mer를 적용했을 때 얻을 수 있는 정보

주어진 문자열에서 등장할 수 있는 길이 k의 문자열 집합을 의미한다. 유전학에선 DNA 염기서열을 일부분 읽어가며 얻을 수 있는 모든 부분 염기 서열을 일컫는다. 가령 [그림 4]와 같은 염기 서열이 주어졌다고 하자. 이때 k=3일때 k-mer를 적용하게 된다면 인접한 3개의 문자를 기준으로 정보를 추출하게 된다. 이렇게 얻어낸 정보를 토대로 분포 계산, 연결관계 파악 등에 활용할 수 있다. 본 과제에선 One-hot vector 방식과 k-mer 방식을 결합하여 인코딩을 수행하고 유의미한 파생변수를 생성할 생각이다.

### > Multi Layer Perceptron (MLP)

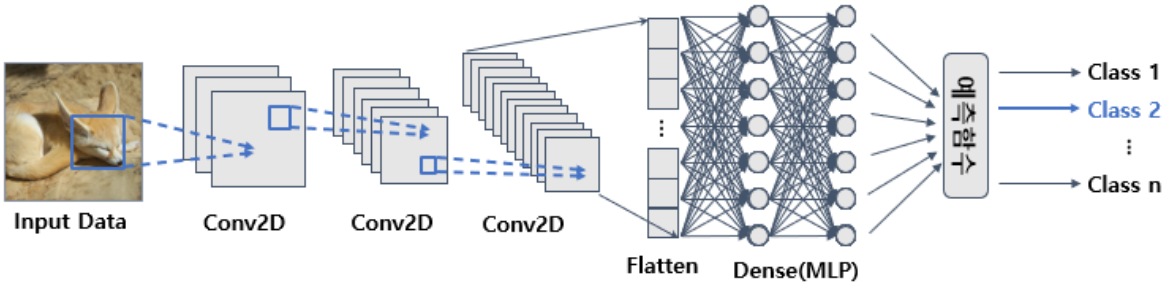



[그림 5] 다층 퍼셉트론의 개형

MLP는 [그림 5]의 형태처럼 단일 퍼셉트론들이 2층 이상 중첩된 fully connected된 구조로, 속성값이 들어가는 Input Layer와 예측 결과를 출력하는 Output Layer, 그리고 실질적인 연산 과정이 이루어지는 Hidden Layer가 존재한다. 각 노드간의 연결 정보에는 개별적인 가중치들이 존재하며, hidden unit의 수는 보통 실험적으로 조절해가며 적정 개수를 파악하게 된다.

은닉층이 단일 형태로 되어있는 3층 구조 MLP를 활용했을 때 Hidden layer의 unit 수가 매우 충분하다면, 어떠한 함수의 형태건 학습을 통해 임의의 함수로 근사시킬 수 있음은 자명하다.[6] 그러나, 단순히 hidden unit의 수만 증가시키면 중간 unit의 수가 너무 많아지고 학습 시간도 과도하게 소요된다. 따라서 본 과제에선 일정 개수의 노드로 이루어진 층을 여려 겹 쌓아서 모델 학습 및 예측을 진행하려 한다.

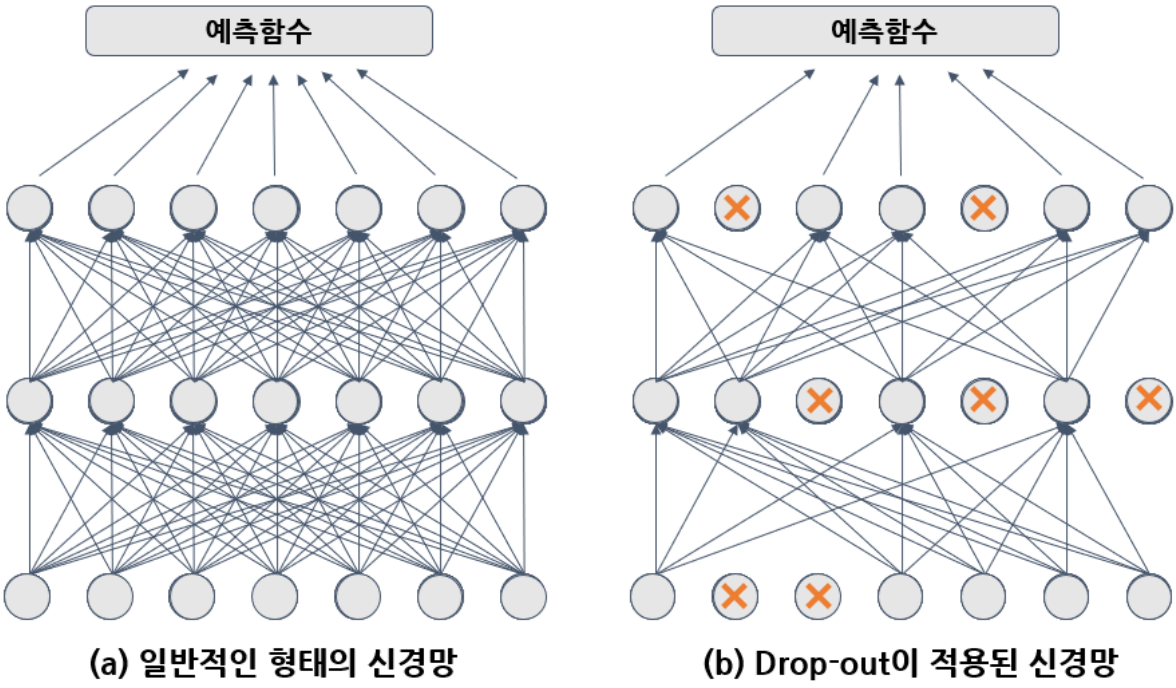
### > Convolutional Neural Network (CNN)




[그림 6] 합성곱 신경망의 개형

CNN는 주로 이미지에서 특징 추출 및 분류 등을 위해 사용하며, [그림 6]의 형태처럼 주어진 이미지에서 특정 영역에 해당 합성곱 결과를 계산하는 Conv2D 레이어, Conv2D 결과를 단층의 형태로 풀어내는 Flatten 레이어, 그리고 일반적인 MLP 레이어와 예측함수 이후의 일반적인 딥러닝 학습 레이어의 형태로 구성된다. 본 과제에서는 Motif의 특징을 추출해내기 위해서 CNN 활용을 시도할 예정이다.

> Drop-out 규제기법




[그림 7] 일반적인 MLP와 Drop-out이 적용된 MLP의 개형

[그림 7(a)]와 같이 MLP를 활용하는 모델의 경우 Tree 모델처럼 Train data에만 과도하게 적합되어 Test data를 예측하기 힘들어지는 Overfitting 현상이 발생한다. 따라서 결정 트리에서 가지치기(Pruning)를 수행하듯, 신경망에서도 확률적으로 각 노드를 학습하지 않는 방식을 사용하게 된다. 이러한 과정을 Drop-out이라고 일컫는다.

Drop-out를 각 레이어 중간중간에 적용하게 되면, [그림 7(b)]처럼 무작위로 노드가 끊어지고, 해당 노드는 현재 학습 과정에서 가중치가 변화하지 않게 된다. 따라서 특정 가중치에만 치우치지 않고 학습이 가능하며, 그로 인해 일반화된다는 장점이 생긴다. 보통 20% ~ 50% 정도로 뉴런을 끊으며, 너무 적은 비율을 사용하면 과대적합이, 너무 많은 비율을 사용하면 과소적합이 발생하므로 적절히 비율을 조절해야 한다.

> TensorFlow & Keras


[그림 8] TensorFlow[7]와 Keras[8] 라이브러리 로고

딥러닝 학습 모델을 제작하기 위한 수단으로 가장 범용적인 라이브러리인 텐서플로우(TensowFlow)와 케라스(Keras)를 파이썬을 기반으로 하여 활용할 예정이다. 텐서플로우는 구글이

개발한 오픈소스 머신러닝 프레임워크이며, 케라스는 텐서플로우를 기반으로 동작하는 머신러닝 라이브러리이다. 통상적으로 빠른 프로토타입 모델을 생성할 때는 케라스에 내장된 모델만으로 학습 모델을 구성하며, 정교한 모델 제작을 할 때는 텐서플로우를 사용하게 된다. 따라서 케라스를 활용해 모델 틀을 구성하고, 세부 조정은 텐서플로우를 활용해 레이어를 적층할 생각이다.

### > Amino Acid Composition(AAC) & Pseudo Amino Acid Composition(PseAAC)

단백질의 구조와 염기 서열의 정보에 따라서 결합 유무를 판별할 수 있도록 염기 서열의 인코딩뿐만 아니라 단백질 구성 형태를 인코딩하는 기법이 필요하다. 이때, 아미노산의 빈도나 결합 구조 정보를 표현하는 기법이 존재하는데, 대표적으로 AAC와 PAAC가 있다.

AAC는 단백질 시퀀스에서 특정 아미노산이 얼마나 출현하는지를 표현한다. 즉, 단순히 전체 개수 대비 특정 아미노산의 개수의 형태로 표현한다. 그러나, 이는 주어진 아미노산의 결합 구조 정보를 무시한다는 단점이 존재한다. 따라서 결합 정보를 고려하여 단백질 유형을 표현하는 PseAAC라는 기법이 2001년 Kuo-Chen Chou에 의해 등장했다.[9]

본 과제에선 특정 단백질과 염기 서열의 결합 정보가 있을 시, 단백질 정보를 AAC와 PseAAC를 활용하여 인코딩을 수행하고, 결합 예측에 도움을 줄 수 있을지 판별할 예정이다.

# 4. 개발 일정 및 담당 역할

## 1. 개발 프로세스 일정

| 구분 | 추진내용 | 추진일정 (월별, 상/하 구분) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 5월 | | 6월 | | 7월 | | 8월 | | 9월 | | 10월 | |
| | | 上 | 下 | 上 | 下 | 上 | 下 | 上 | 下 | 上 | 下 | 上 | 下 |
| 계획 | 착수보고서 작성 | | | | | | | | | | | | |
| 분석 | 생물학 전문지식 이해 | | | | | | | | | | | | |
| | 사용할 주요 기술 사전 조사 | | | | | | | | | | | | |
| 설계 | 학습 모델 기법 연구 | | | | | | | | | | | | |
| | 데이터셋 소스 탐색 및 선정 | | | | | | | | | | | | |
| | 개발 환경 구축 | | | | | | | | | | | | |
| 개발 | 생성, 판별용 베이스모델 구축 | | | | | | | | | | | | |
| | 학습용 데이터셋 전처리 | | | | | | | | | | | | |
| | 중간보고서 작성 | | | | | | | | | | | | |
| | 모델 학습 및 클리닝 | | | | | | | | | | | | |
| 테스트 | 모델 테스트 및 최적화 | | | | | | | | | | | | |
| 마무리 | 최종보고서 작성 및 발표 | | | | | | | | | | | | |
| | 결과물 업로드 및 후속 처리 | | | | | | | | | | | | |

[표 1] 전체 개발 일정에 대한 Gantt chart

## 2. 조원별 담당 역할

| 구분 | 조원 | 담당업무 |
|---|---|---|
| 1 | 김유진 | DeepBind 모델 연구 및 응용, 학습 모델 구축, 염기 서열 생성자 모델 연구 |
| 2 | 유경민 | AptaNet 모델 연구 및 응용, 학습 모델 구축, 염기 서열 판별자 모델 연구 |
| 3 | 허수민 | 데이터 전처리, 평가 지표 선정, 프로그램 개발 환경 선정 및 구축 |
| - | 전체 | 가용 데이터셋 조사, 모델 테스트 및 최적화, 보고서 및 발표 자료 작성 |

[표 2] 조원별 담당 역할

# 출처 및 참조

[1] Shuanghui Yang, Huan Li, Ling Xu, Zhenhan Deng, Wei Han, Yanting Liu, Wenqi Jiang, Youli Zu, Oligonucleotide Aptamer-Mediated Precision Therapy of Hematological Malignancies, Molecular Therapy - Nucleic Acids, Volume 13, Pages 164-175, 2018.
[2] Yi Xi Wu, Young Jik Kwon, Aptamers: The “evolution” of SELEX, Methods, Volume 106, Pages 21-28, 2016.
[3] Emami, N., Ferdousi, R. AptaNet as a deep learning approach for aptamer–protein interaction prediction. Sci Rep 11, 6074, 2021.
[4] Alipanahi, B., Delong, A., Weirauch, M. et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 33, 831–838, 2015.
[5] Lee G, Jang GH, Kang HY, Song G. Predicting aptamer sequences that interact with target proteins using an aptamer-protein interaction classifier and a Monte Carlo tree search approach. PLoS One. 2021.
[6] K. Hornik, M. Stincombe, and H. White, "Multilayer feedforward networks are universal approximators," Neural Networks, Vol.2, pp.359-366, 1989.
[7] https://www.tensorflow.org 메인 홈페이지 로고
[8] https://keras.io 메인 홈페이지 로고
[9] Ding, Y. S., Zhang, T. L. & Chou, K. C. Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Pept. Lett. 14, 811–815, 2007.