

# 09

# 신약 개발 후보 물질 추천을 위한 머신러닝 모델 설계

소속 정보컴퓨터공학부

분과 A

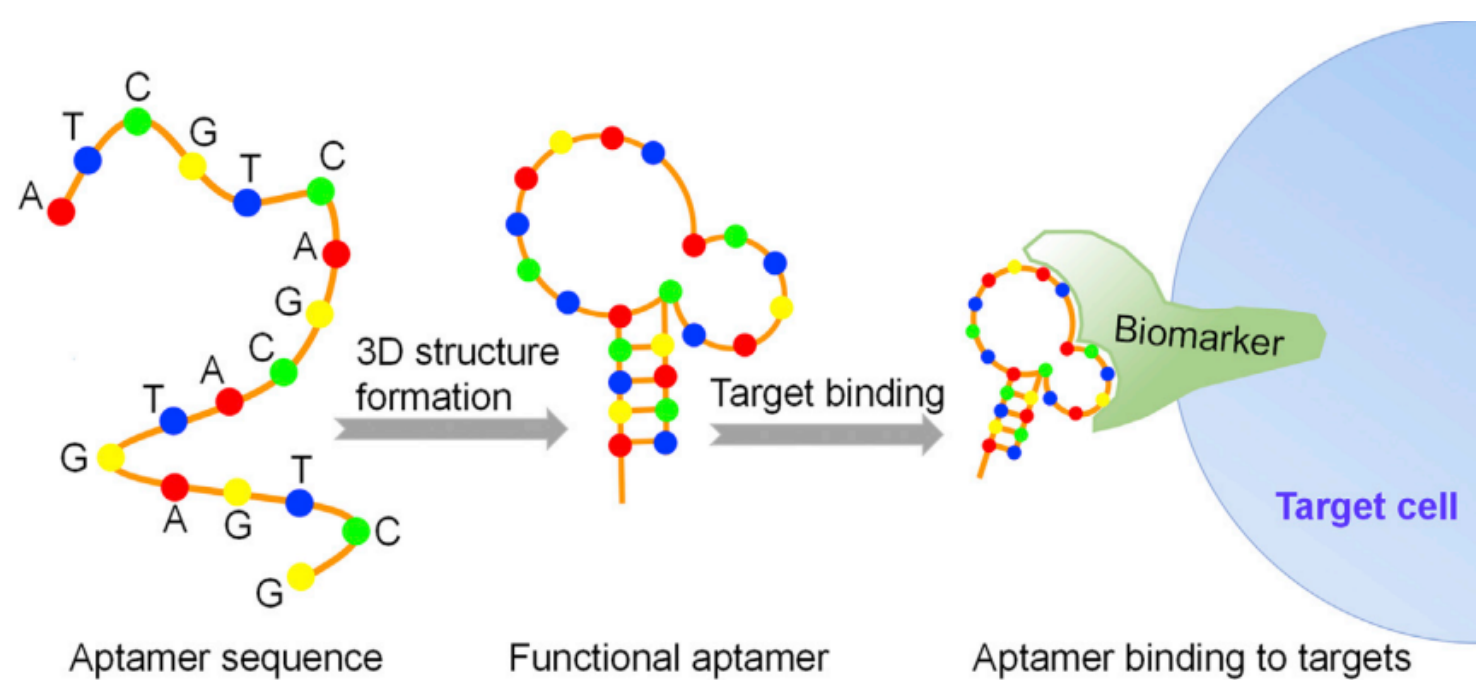
팀명 ML궤는 컴공

참여학생 허수민, 유경민

지도교수 송길태

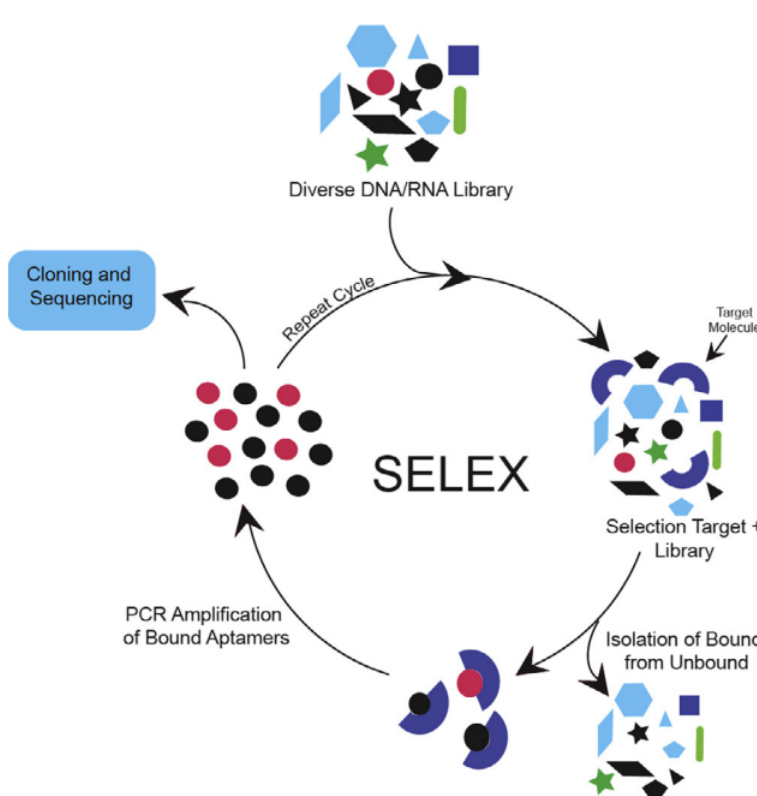
## 개발 배경

### 압타머(Aptamer)란?



- RNA / DNA와 유사한 구조를 가지는 핵산 물질
- 항체와 비교될 수 있을 정도의 높은 결합 친화도
- 표적 단백질에 다양한 형태로 결합 가능
- 암 치료제 분야에서 사용

### SELEX 란?



- 다양한 후보의 압타머 중 결합 친화도가 높은 물질을 선별하는 대표적인 과정
- 유전 물질을 합성 & 분리하는 작업을 반복적으로 수행
- 시간이 오래 소요

➡ 알고리즘 기반 예측 모델 등장

## 개발 내용

### 데이터 전처리

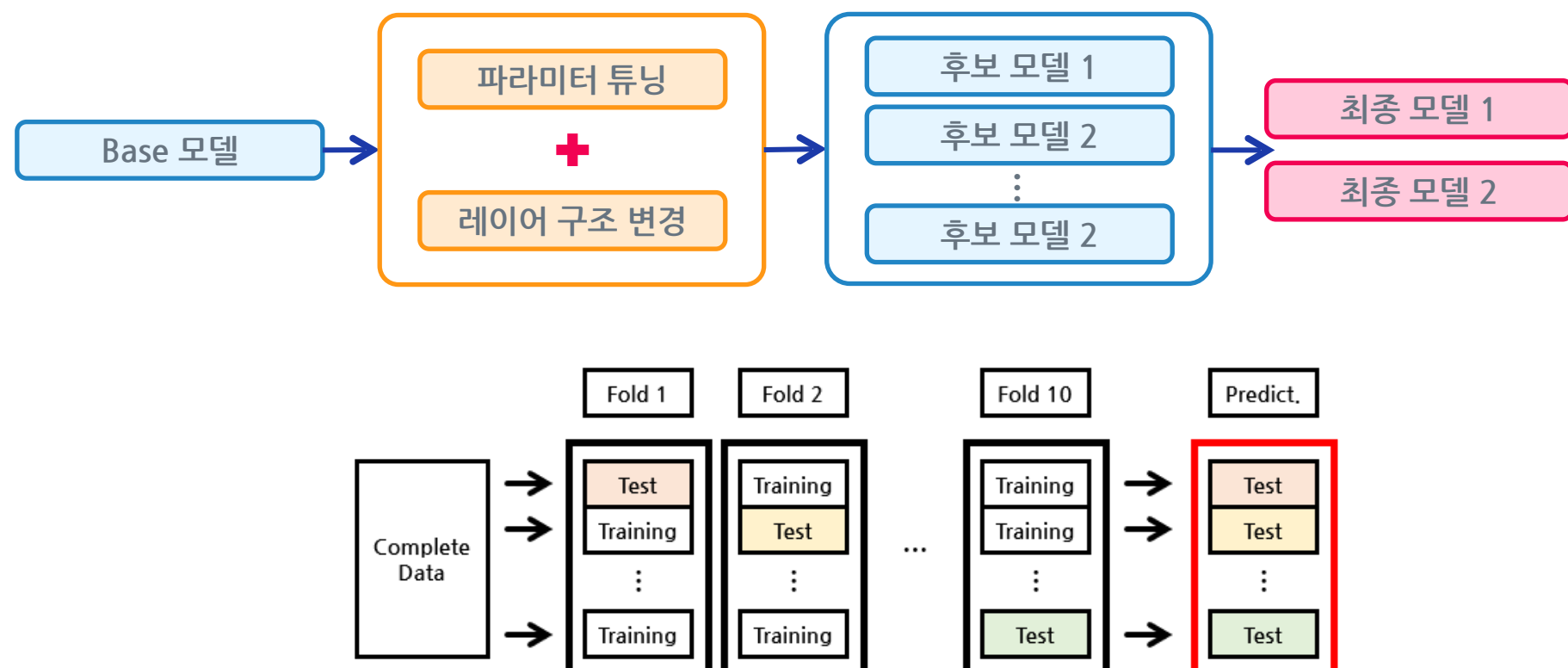
염기서열	T	A	A	C	T	G	A	A	C	T	G
3-mer	T	A	A		G	A	A				
		A	A	C		A	A	C			
			A	C	T		A	C	C		
				C	T	G		C	C	T	
					T	G	A		C	T	G
메타정보	TAA: 1	AAC: 2	ACT: 1	CTG: 2							
	TGA: 1	GAA: 1	ACC: 1	CCT: 1							

Rcpi (version 1.8.0)	
extractProtPAAC: Pseudo Amino Acid Composition Descriptor	
Description	Pseudo Amino Acid Composition Descriptor
Usage	<code>extractProtPAAC(x, props = c("hydrophobicity", "hydropathicity", "isoelectricity"), lambda = 30, w = 0.45, cutprop = NULL)</code>
Arguments	<p><code>x</code>: A character vector, as the input protein sequence.</p> <p><code>props</code>: A character vector, specifying the properties used. 5 properties are used by default, as listed below:</p> <ul style="list-style-type: none"> <li>"hydrophobicity": Hydrophobicity value of the 20 amino acids</li> <li>"hydropathicity": Hydropathicity value of the 20 amino acids</li> </ul>

- k = 1,2,3,4인 경우의 메타정보 추출 (k-mer 지표)
- 각 k값을 기준으로 상대적인 비율(합 1)의 형태로 염기서열 전처리 수행
- AAC기법과 PseAAC 기법을 이용해 단백질 구성 형태를 고려하여 일반적인 모델 제작
- k-mer 지표와 PseAAC 정보를 합쳐서 각 API별 학습용 데이터 구축

PseAAC: 2001년 Kuo-Chen Chou에 의해 등장한 결합 정보를 고려한 단백질 표현 기법

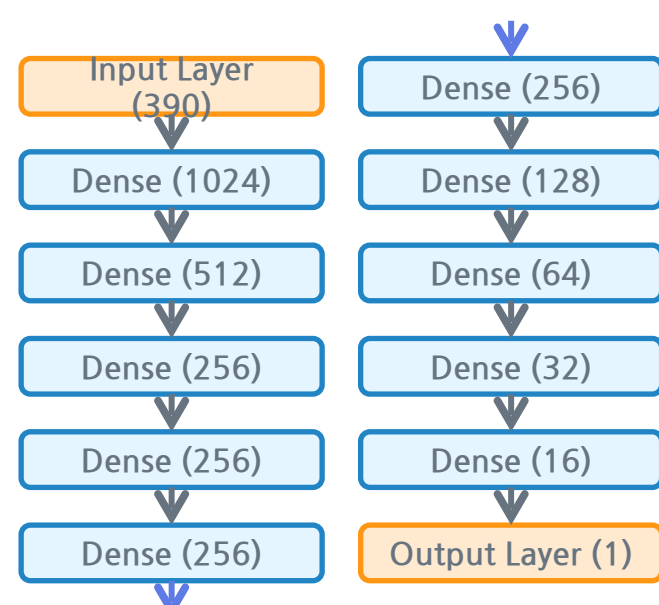
### 모델 구축



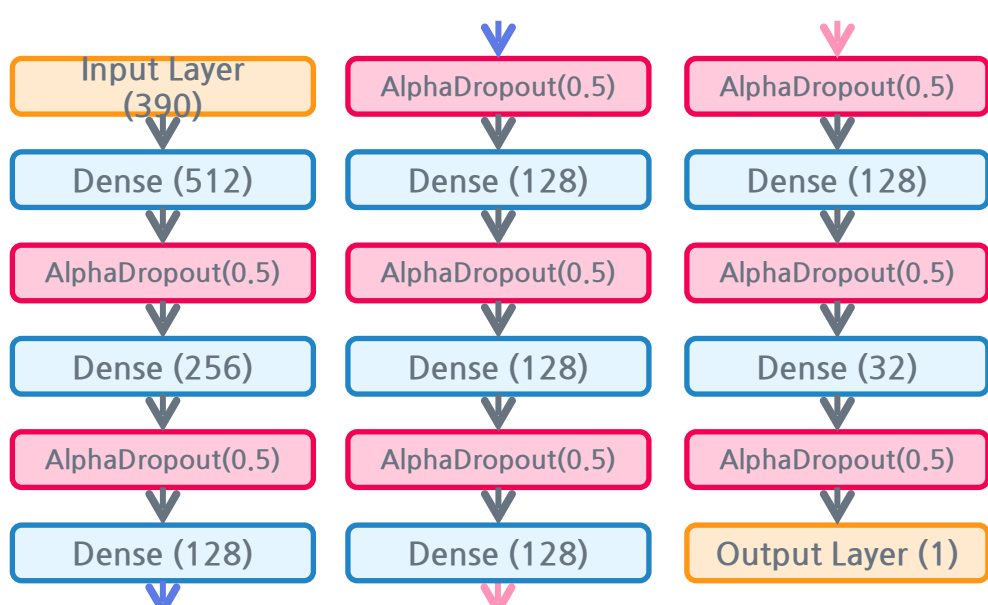
- 구조적으로 변형이 용이한 AptaNet 모델을 기반으로 튜닝 수행
- Grid Search를 이용해 최적의 batch\_size와 epochs, Learning Rate, Rho 값을 찾음
- 10-fold validation을 이용해 객관적인 성능 측정
- 불균형 데이터에도 효과적인 성능 지표인 F1-score를 기준으로 성능 비교
- 약 20가지의 Layer 변형 모델을 비교해 가장 성능이 높은 2가지 모델 선정

## 연구 성과

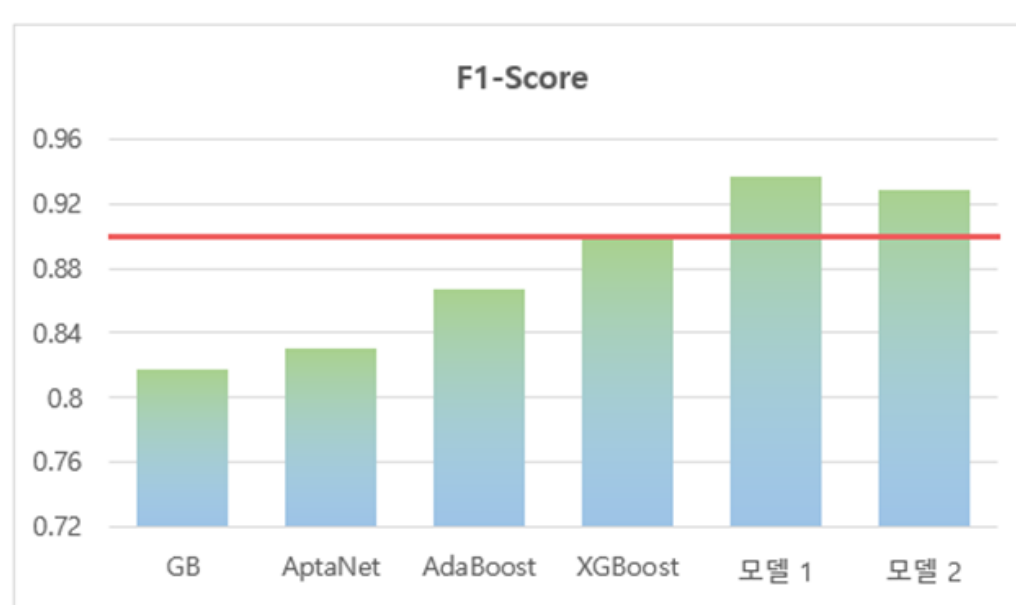
### 모델1 구조



### 모델2 구조



### 기존 여러 모델들과 성능 비교



- 기존 모델들은 0.9 미만의 성능을 보임
- 모델 1의 F1-Score : 0.937
- 모델 2의 F1-Score : 0.929



기존 모델 중에 가장 성능이 좋은 XGBoost에 비해 약 3~4% 성능 개선