JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 27 No. 8, pp. 49-59, August 2022
https://doi.org/10.9708/jksci.2022.27.08.049

Small CNN-RNN Engraft Model Study for Sequence Pattern Extraction in Protein Function Prediction Problems

Jeung Min Lee*, Hyun Lee**

*Student, Bio Big Data Convergence Major, Dept. of Computer and Electronics Convergence Engineering, Sunmoon University, Asan, Korea
**Professor, Division of Computer Science and Engineering, Sunmoon University, Asan, Korea

[Abstract]

In this paper, we designed a new enzyme function prediction model PSCREM based on a study that compared and evaluated CNN and LSTM/GRU models, which are the most widely used deep learning models in the field of predicting functions and structures using protein sequences in 2020, under the same conditions. Sequence evolution information was used to preserve detailed patterns which would miss in CNN convolution, and the relationship information between amino acids with functional significance was extracted through overlapping RNNs. It was referenced to feature map production. The RNN family of algorithms used in small CNN-RNN models are LSTM algorithms and GRU algorithms, which are usually stacked two to three times over 100 units, but in this paper, small RNNs consisting of 10 and 20 units are overlapped. The model used the PSSM profile, which is transformed from protein sequence data. The experiment proved 86.4% the performance for the problem of predicting the main classes of enzyme number, and it was confirmed that the performance was 84.4% accurate up to the sub-sub classes of enzyme number. Thus, PSCREM better identifies unique patterns related to protein function through overlapped RNN, and Overlapped RNN is proposed as a novel methodology for protein function and structure prediction extraction.

▸ **Key words**: PSSM, Deep learning, Protein Function Prediction, Feature Engraft Model, Overlapped RNN

[요 약]

본 논문에서는 2020년 기준 단백질 서열을 이용한 기능과 구조 예측 분야에서 가장 많이 사용되고 있는 딥러닝 모델인 CNN과 LSTM/GRU 모델을 동일한 조건 하에 비교 평가한 연구를 토대로 새로운 효소 기능 예측 모델인 PSCREM을 설계하였다. CNN 합성곱 시 누락되는 세부 패턴을 보존하기 위하여 서열 진화정보를 이용하였으며 중첩 RNN을 통해 기능적으로 중요한 의미를 가지는 아미노산 간의 관계 정보를 추출하고 특징 맵 제작에 참조하였다. 사용된 RNN 계열의 알고리즘은 LSTM과 GRU로 보통 stacked RNN 기법으로 100 units 이상 2~3회 쌓는 것이 일반적이나 본 논문에서는 10, 20 unit으로 구성한 뒤 중첩시켜서 특징 맵 제작에 사용하였다. 모델에 들어가는 데이터는 단백질 서열 데이터로 PSSM profile로 가공한 뒤 사용되었다. 실험 결과 효소 번호 첫 번째 자리를 예측하는 문제에 대해 86.4%의 정확도를 나타냄을 입증하였고, 효소 번호 3번째 자리까지 예측 정확도 84.4%의 성능을 내는 것을 확인하였다. PSCREM은 Overlapped RNN을 통해 단백질 기능에 관련된 고유 패턴을 더 잘 파악하며 Overlapped RNN은 단백질 기능 및 구조 예측 추출 분야에 새로운 방법론으로서 제안된다.

▸ **주제어**: PSSM, 딥러닝, 단백질 기능 예측, 특징 접목 모델, 중첩 RNN

• First Author: Jeung Min Lee, Corresponding Author: Hyun Lee
*Jeung Min Lee (starleejeung@gmail.com), Bio Big Data Convergence Major, Dept. of Computer and Electronics Convergence Engineering, Sunmoon University
**Hyun Lee (mahyun91@sunmoon.ac.kr), Division of Computer Science and Engineering, Sunmoon University
• Received: 2022. 07. 25, Revised: 2022. 08. 22, Accepted: 2022. 08. 23.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

단백질은 20종류의 아미노산으로 구성되는 생체 고분자의 일종으로 대부분의 생명 활동에 관여한다. 단백질의 구조와 기능은 아미노산 서열이 내포하는 생명 정보를 따라 결정되므로 일반적으로 단백질의 서열이 흡사할수록 그 구조와 기능 또한 유사하다. 그러나 같은 기능을 가졌음에도 단백질 서열의 유사도가 낮은 경우와 완전히 다른 경우 역시 존재한다. 생명 과학에서 기능적, 생물학적으로 어떠한 의미를 가진다고 추측되는 단백질 서열의 특정한 부분들을 모티프(Motif) 또는 도메인(Domain)이라고 부르는데 이 특이지점이 다른 기능을 가진 단백질 사이에서 공통으로 발견되기도 한다. 이렇듯 단백질은 변이(Mutation)가 심하며 이 때문에 특정 조합의 아미노산 서열이 기능과 구조에 관련 있는 것은 명백하나 정확히 어떤 패턴이 단백질의 구조와 기능을 결정짓는지는 아직까지도 상세히 밝혀지지 않았다.

이를 극복하기 위해 많은 연구자들이 단백질 서열 내 고유한 패턴을 찾는 시도를 해왔다. 그중 서열의 진화정보를 이용해 단백질의 기능과 구조를 예측하는 방식이 가장 많이 시도되었다. 단백질은 어떤 생물학적 의미를 가지면서 아주 흡사한 서열임에도 부분적 변이에 따라 세부 서열 표현이 일정 부분 차이 날 수 있다. 즉, 특정 위치에 해당하는 아미노산이 다른 아미노산으로도 대체되어도 기능을 유지하는 경우가 존재한다. 이에 우연히 비슷할 수 있는 경우의 수를 배제하고 단백질 서열 내에 공통적으로 출현하는 아미노산의 교차성을 점수화하여 나타낸 계산법이 등장하였다. 이것을 위치 특이적 득점 행렬(Position Specific Scoring Matrix; PSSM)이라고 부른다. PSSM은 행렬 안에 단백질 서열의 진화정보를 담고 있으며 이러한 특성 탓에 지금도 단백질 구조 예측 문제에서 자주 쓰이고 있다.

그림 1로 요약한 것처럼 서열의 진화정보를 이용하는 연구로 Y. Liang et al.(2015), POSSUM(2017) 과 같이 구조를 예측하는 연구[1,2], Mousavian Z(2016)과 같이 약물 표적 상호 작용을 예측하는 연구[3], SNARE-CNN(2019)와 같이 SNARE 단백질을 예측하는 연구[4]가 있으며, EPTool(2021), SNB-PSSM(2021)과 같이 서열 진화정보 매트릭스 자체를 더 개선시키고자 하는 연구 또한 활발하다[5,6]. 그뿐만 아니라 ECPred(2018), EnzyNet(2018), MF-EFP(2020), UDSMProt(2020)과 같이 서열 진화정보를 활용해 단백질 기능을 예측하기 위한 연구 또한 다수 수행되었다[7-10]. 본 연구 또한 서열의 진화정보를 이용해 단백질 기능과 관련된 고유 패턴을 추출하기 위한 딥러닝 모델을 설계하고 실험하였다.

본 논문에서 제안하는 것은 선행된 비교 실험[11]의 결과를 기반으로 구성한 CNN-RNN 특징 접목 모델이다. 입력으로는 PSSM profile로 변환된 서열 데이터를 사용하였다. 모델에는 CNN과 LSTM, GRU가 모두 사용되었으며 RNN 계열인 LSTM과 GRU 모델은 일반적인 적용법인 쌓기가 아니라 중첩되었다. 중첩 CNN과 중첩 RNN으로 추출한 특징값을 길게 접목하여 새로운 특징 맵을 구성하였다. 서열 검색 도구를 이용해 만들어진 서열 진화정보가 모델의 입력 데이터로 사용되며 모델은 최종결과물로 효소 번호를 산출한다.

본 논문에서는 먼저 아미노산 서열의 텍스트 데이터로 모델 자체의 패턴 추출 성능을 검증한 뒤에 PSSM profile을 적용한 목적의 모델을 설계하였다. 이에 제안하는 모델의 성능 검증 실험은 총 4가지로 구성되었다. 서열 진화정보로 변환하지 않고 단백질 문자열 자체를 사용하여 효소 번호 첫 번째 자리까지 예측하는 실험, 서열 진화정보를 이용하여 효소 번호 첫 번째 자리를 예측하는 실험, 같은 조건으로 서열 진화정보를 사용하여 효소 번호 세 번째 자리까지 예측하는 실험, 제안 모델과 다른 모델을 동일한 데이터로 비교하는 실험으로 이루어진다.

본 논문의 구성은 다음과 같다. 2장에서 주요 배경지식과 선행 연구에 관해 설명한다. 3장에서 제안 모델과 실험 방법을 소개한다. 4장에서 제안 모델 성능 검증 실험 결과를 분석한다. 5장에서 결론을 맺고 마무리한다.

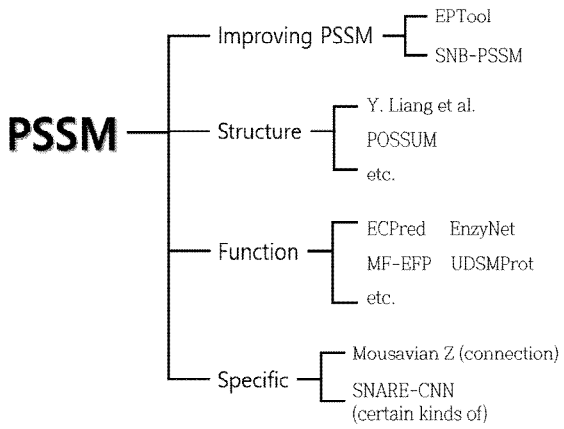

Fig. 1. Summary tree about the study using PSSM

II. BackGround

1. Enzyme in Protein Function Prediction

효소란 특수한 기능을 가진 단백질로 자기 자신은 변하지 않으면서 대부분의 생명 활동과 대사 활동 과정을 촉매한다. 모든 단백질이 효소는 아니지만 모든 효소는 단백질로 이루어진다. 효소는 자물쇠-열쇠 모델이라고 불리는 특성이 있으며 활성 부위의 형태에 따라 결합되는 기질이 결정된다[12]. 열 또는 pH 변화로 단백질이 손상되어 본래의 구조를 잃으면 효소가 가진 본래의 기능 또한 잃어버리므로 효소의 구조는 효소의 기능과 밀접한 연관성을 가진다[13].

국제 생화학 연합 효소 위원회는 효소의 기능을 마침표로 구분된 4자리의 숫자로 표기하였다. 각 자리의 숫자는 특정 단백질의 반응식과 결합되는 기질을 의미하며 계층적으로 구성되는 네 자리 숫자의 조합은 효소가 가진 특수한 기능을 나타낸다. 이와 같이 특수한 기능의 단백질을 역할에 따라 정리한 효소 번호 데이터는 변형이 다양한 단백질의 특성상 종종 불완전한 결과를 나타내기도 하나[14], 체계적으로 정립된 역사가 길고 데이터의 신뢰도가 높아[15] 딥러닝을 이용한 단백질 기능 예측 연구에 Label로써 사용되기 적합하다.

2. Position Specific Scoring Matrix(PSSM)

위치별 득점 매트릭스, 위치 특이적 득점 행렬(PSSM)이라고 불리는 이 매트릭스는 단백질 서열의 진화정보를 보존한다. 그 때문에 주로 서열을 사용한 단백질 구조 예측 연구에 자주 사용된다.

BLAST(Basic Local Alignment Search Tool)는 데이터베이스의 다른 서열과 비교하여 로컬 정렬 영역을 식별하고 주어진 점수를 초과하는 정렬을 표시하는 서열 유사성 검색 방법이다. PSI-BLAST(Position-Specific Iterative) 또한 이 중 하나이며 기존 검색을 정확도 측면에서 좀 더 개선한 것이다[16]. 본 논문에서 사용된 PSSM은 PSI-BLAST의 수행 결과로서 만들어졌다.

PSSM 매트릭스는 식 (1)과 같은 L × 20의 행렬로 구성된다. L 행은 단백질 서열의 길이를 나타내고, 20열은 20개의 아미노산을 나타내고 있으며, 각 아미노산이 다른 아미노산으로 돌연변이 될 확률을 나타낸다. 20 열은 순서대로 A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V로 각 알파벳은 고유한 하나의 아미노산을 지정한다.

$$PSSM = \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,20} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ P_{L,1} & P_{L,2} & \cdots & P_{L,20} \end{bmatrix}_{L \times 20} \quad (1)$$

매트릭스 안의 실수들은 단백질 서열의 진화정보를 점수로 나타낸다. BLOSUM62 점수행렬의 표기를 따라 한 아미노산이 다른 아미노산으로 바뀔 가능성이 큰 경우는 양수로 표기하고, 두 아미노산이 서로 잘 바뀌지 않을수록 낮은 음수로 표현된다. 0의 값은 특별한 의미 없이 우연히 바뀌는 경우를 의미한다[17].

3. Previous Work: Model Comparison

모델에 적용된 파라미터는 본 논문에 앞서 수반된 모델 비교 실험[11] 결과를 기반으로 구성되었다. 선행 연구는 2020년 기준 단백질 서열을 다루어 단백질의 기능이나 구조를 예측하는 융합 분야에서 자주 보이는 딥러닝 모델인 CNN, LSTM, GRU의 단일 모델 성능과 CNN-LSTM, CNN-GRU의 결합 모델의 성능을 동일한 조건 하에 비교 실험하였다.

이를 위해 비교실험 전 CNN에 사용될 필터 사이즈 선별 실험과 RNN 계열 모델에 사용될 히든 유닛과 깊이 선정 실험이 선행되었으며, 중첩 CNN을 사용하는 경우 필터 크기는 작은 것을 여러 번 중첩하는 것이 가장 성능이 좋았으며, RNN은 stacked 횟수가 많아질수록 히든 유닛의 개수와는 상관없이 성능이 떨어지는 것을 확인하였다.

해당 실험에서 가장 결과가 좋았던 알고리즘은 50 unit을 2회 쌓은 LSTM 알고리즘이었으며 이는 단백질 서열이 명백히 순서를 지니며, 순서에 따른 의미를 가진 정보로써 시계열 데이터 처리에 적합한 알고리즘이 패턴 처리에 유리함을 나타내었다.

그러나 LSTM의 경량형이자 비슷한 성능을 가진다고 보고 된 GRU는 해당 논문에서 수행된 모든 실험에서 LSTM과 분명한 성능 차이를 보였다. 단일 모델 성능에서 항상 2.2% 이상의 정확도 차이를 보였던 두 RNN 계열의 성능 차는 CNN이 앞에 결합되자 1.7%로 그 성능 차가 줄어들었는데 이는 GRU 자체의 파라미터 문제나 내부 계산법의 문제가 아닌 CNN으로 인해 제거되는 서열의 디테일한 표현에 의한 두 RNN 계열 모델의 단어 간의 관계성 파악이 상이했던 것으로 이해된다.

III. PSCREM Method

본 논문에서 제안하는 모델인 PSCREM은 Y. Kim(2014)[18]에서 수행한 문장 분류를 위한 컨볼루션 신경망 구조인 중첩 CNN의 특징 맵에 적은 수의 Unit을 적용한 중첩 LSTM과 GRU의 결과를 더하여 특징 맵을 확장한다. 중첩 RNN은 Overlapped CNN이 특징 맵을 제작하는 방식과 같이 RNN 계열의 알고리즘을 병렬로 계산한 후 도출된 벡터를 합치하여 활용한다. 중첩 CNN 모델은 단백질 기능 예측 페이퍼인 DeepEC[19]에서도 사용되었으나, LSTM과 GRU는 이와 같이 중첩 RNN으로 사용된 적이 없다. 일반적으로 RNN 계열 알고리즘은 stacked RNN 기법으로 100 units 이상 적용한 층을 2~3 깊이로 쌓아서 사용되어지나 본 논문에서는 기존의 10분의 1 수준인 10, 20 units의 RNN 층이 각각 단층으로 적용되었다. 효소 번호 예측을 위한 모델의 전체적인 데이터 처리 흐름과 구조는 그림 2와 같다.

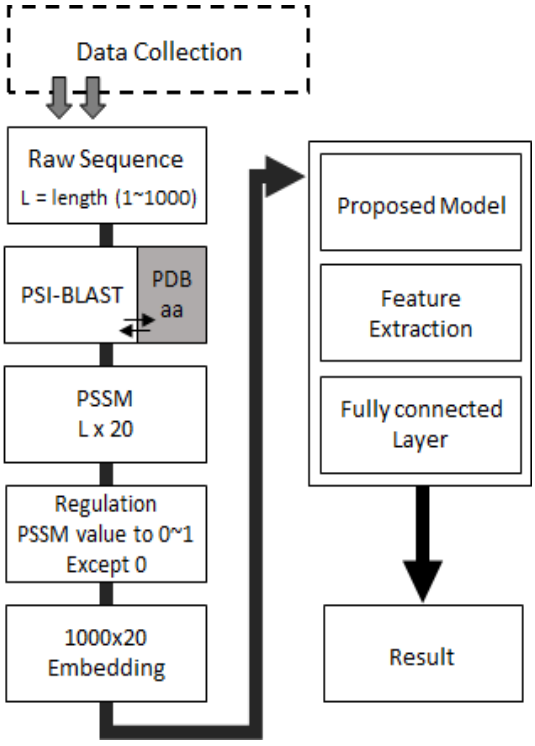

Fig. 2. The work procedure of the proposed method for EC number prediction

3.1 Data Preprocessing

3.1.1 Data Collection

실험에 사용될 단백질 시퀀스는 UniProtKB 2022_01 Swiss-Prot을 사용하였다. 하나의 서열에 각기 다른 효소 번호를 여러 개 가진 효소는 데이터 수집 과정에서 전부 제외하였다. 또한, 효소 번호 중분류와 소분류에 90번대가 포함된 서열도 제외하였다. 효소 번호가 90번 대에 배정된 경우는 기존의 분류에 속하면서 정확히 구분하기 어려운 경우를 몰아서 배정하는 나머지 번호이기 때문에 패턴 파악에 혼란을 줄 수 있기 때문이다.

데이터 수집 과정을 통해 수집된 총 시퀀스 개수는 566,996개였으며 데이터 선별과정을 통해서 최종적으로 남은 시퀀스 개수는 237,923개다.

3.1.2 Converting to PSSM

Ruibo Gao et al.(2019)[20]의 연구에서는 단백질의 구조 예측에 사용되는 데이터 처리 방식인 PSSM profile 정보를 사용해 효소의 기능을 예측하였다. 본 논문에서 또한 단백질 구조 예측에 자주 사용되는 데이터 처리방식인 PSSM proflie을 사용하여 데이터를 전처리하고 모델에 적용한다.

보통 서열 검색 시 사용되곤 하는 nr 데이터베이스가 아니라 PDBaa 데이터베이스를 사용한 이유는 다음과 같다. PDB(Protein Data Bank)는 Secondary Database로서 NCBI와 같은 Primary Database(아카이브 데이터베이스)에서 특정 목적을 가지고 가공된 서열 데이터가 모이는 장소이다. 그 중 PDB는 단백질의 구조정보를 서열과 연관지어 집중적으로 등록하고 있으며 이때 올라오는 정보들은 전부 사람이 직접 실험하고 증빙한 정보만이 올라온다. 따라서 여타 2차 데이터베이스보다 가지고 있는 서열 정보의 수는 적지만 구조정보에 관해서라면 가장 신뢰할 수 있는 데이터베이스이다.

본 논문은 구조 예측에 사용되는 PSSM profile 정보를 이용해 효소의 기능을 예측한다. 이에 구조적으로 확실히 검증된 시퀀스만을 사용하여 서열의 상동성과 그 사이의 특이 패턴을 파악할 필요성이 있었다. 따라서 구조적으로 검증되지 않은 서열까지 포함된 nr 데이터베이스가 아닌 PDBaa Database를 서열 검색의 지정 데이터베이스로 선택하였다.

각 서열의 PSSM은 0.001의 E-값으로 3번 반복되었고, PDB 데이터베이스에 대체 PSI-BLAST[16]를 실시하였다. 이를 통해 도출된 매트릭스의 형태는 그림 3과 같다.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1 M	-1	-1	-2	-3	-2	0	-2	-3	-2	1	2	-1	6	0	-3	-2	-1	-2	-1	1
2 P	-1	-2	-2	-2	-3	-1	-1	-2	-2	-3	-3	-1	-3	-4	8	-1	-1	-4	-3	-3
3 E	-1	1	0	1	-4	2	4	-2	0	-3	-3	3	-2	-3	-1	0	-1	-3	-2	-3
4 R	-2	5	0	-2	-4	1	0	-2	5	-3	-3	1	-2	-2	-2	-1	-1	-3	0	-3
5 L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
6 Q	-1	1	0	1	-3	5	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-2	-2	-2
7 V	0	-3	-3	-3	-1	-2	-3	-3	-3	3	1	-2	1	-1	-3	-2	0	-3	-1	4
8 Y	-2	-2	-2	-3	-3	-2	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
9 K	-1	2	0	-1	-3	1	1	-2	-1	-3	-3	5	-1	-3	-1	0	-1	-3	-2	-2
10 C	0	-4	-3	-4	9	-3	-4	-3	-3	-1	-1	-3	-2	-3	-3	-1	-1	-3	-3	-1

Fig. 3. Ascii PSSM

그러나 오리지널 PSSM을 사용할 시 음수 값이 포함되므로 컴퓨터가 계산을 용이하게 해주기 위해 각 값을 0-1 사이로 만들어 주었다. 이때 아래 해당하는 수식을 사용해서 전체 값을 0-1 사이의 값으로 평준화하였다.

이때 사용한 방정식은 다음 식 (2)와 같다.

$$p'_{i,j} = \frac{p_{i,j} - p_{\min}}{p_{\max} - p_{\min}} \ (i = 1,...,L;\ j = 1,...,20) \quad (2)$$

이 식은 0을 제외한 숫자들에만 적용되었다. 0을 제외하지 않으면 아래 임베딩에서 서술할 zero-padding에 의하여 서열에 존재하지 않는 정보가 서열상의 우연 정보로 인식될 가능성이 있었다. 하여 특별한 의미가 있지 않은 0을 제외하고 음수와 양수 값만 보존하였다.

3.1.3 Embedding

순서 정보를 유지하기 위한 단순 임베딩 방식을 선택하였다. PSI-BLAST를 통해 얻은 Ascii PSSM의 행은 서열의 길이다. 그러나 단백질 서열의 길이가 모두 제각각이므로 모델에 입력으로 들어가기 위해서는 공통된 크기로 맞춰야 할 필요성이 있었다. 실험에 사용된 서열은 전부 길이 1000 이하의 서열들만 사용되었고 순서 정보가 보존될 수 있도록 PSSM을 20 × 20으로 요약하는 일 없이 모델에 들어가는 모든 매트릭스의 크기를 1000 × 20으로 바꾸어 주었다. 시퀀스의 길이가 1,000보다 짧은 경우 나머지 행렬에 전부 0을 채워 넣었다.

매트릭스에 0을 패딩(padding) 하는 방법에 있어 Lopez-del Rio et al.[21]은 dense padding의 경우 0을 매트릭스의 전후 어디에 채워도 모델의 성능에 차이를 보이지 않았다고 보고하였으므로 일반적으로 사용되던 Post-zero-padding 방식으로 결정되었다.

3.2 Datasets

본 연구에서 사용된 데이터 세트는 총 3가지이다. 비교 실험에서 사용된 다른 모델과의 성능 비교를 위한 모델 검증용 일반 텍스트 데이터 세트 한 개와 PSSM profile로서 본 연구의 목적에 맞게 구성된 데이터 세트 두 개로 이루어진다. 모든 데이터는 매 실험 시작 전에 항상 Train, Valid, Test 6:2:2의 비율로 나뉘었다. 표 1에 데이터세트 정보를 요약하였다.

Table 1. Summarized Dataset Informations

Name	Main class Text Dataset	Main class PSSM Dataset	Sub-sub class PSSM Dataset
# of sample	70,000	70,000	237,973
Input type	Text Sequence	Ascii PSSM	Ascii PSSM
Target class	Main	Main	Sub-sub
# of class	7	7	139
Dataset type	Balanced	Balanced	Extreme Imbalanced

3.2.1 Main class Text Dataset

총 데이터 개수는 70,000개이다. EC 번호 대분류 7가지를 분류한다. 데이터 세트는 균형하며 각 class마다 10,000개로 모두 동일한 개수의 데이터가 적용되었다. 시퀀스가 PSSM으로 가공되지 않았으며 고유한 문자 서열을 그대로 가지고 있다. 이 데이터는 PSSM과 상관없이 새로 설계된 모델의 자체의 패턴 파악 능력을 검증하기 위해 사용되었다.

3.2.2 PSSM Dataset

모든 데이터에 대해 PSI-BLAST를 먼저 수행하였고 그렇게 얻은 2차원 배열 형식의 ascii_pssm_output을 flatten 시켜 1차원 리스트 형식으로 저장하였다. 이 문자열 List는 모델에 특징 추출 층에 들어가기 직전에 array 형태로 바뀐 뒤 1000 × 20 형태로 reshape 되어 사용되었다. PSSM profile화 이후 기존의 텍스트 데이터는 약 6,200배 증가하였으며 90MB의 서열 데이터는 총 58.7GB로 증가하였다.

지도학습에 필요한 Main class label, Sub-Sub class 라벨에는 One-hot encoding을 수행하였다. 다만 label의 경우 변환하는데 걸리는 시간이 짧아 미리 가공해서 저장하지 않았고 모델에 들어가기 직전에 변환되어 학습에 사용되었다. 가공된 데이터 세트의 정보는 다음과 같다.

3.2.2.1 Main class Dataset

총 데이터 개수는 70,000개이다. EC 번호 대분류 7가지를 분류한다. 모든 데이터는 1만 개씩 고르게 분배되었다. 시퀀스가 PSI-BLAST 결과로 구성된 PSSM이 flatten 되어 한 줄의 리스트 문자열로 적재되어있다는 점만 제외하고는 Main class Text Dataset와 동일하다.

3.2.2.2 Sub-Sub class Dataset

총 데이터 개수는 237,973개이다. 라벨 개수는 139개이다. 데이터 세트는 극심하게 불균형하다. 가장 많은 데이터는 약 1만 개고 가장 적은 데이터는 약 102개이다. Main class PSSM Dataset과 똑같이 시퀀스가 PSI-BLAST 결과로 구성된 PSSM이 한 줄의 문자열 List 형태로 적재되어있다. Train 141,363개, Valid 47,293개, Test 47,293개로 구성되고 실험마다 50%씩 각 데이터 세트에서 제각기 표집된 부분 데이터 세트가 매 실험마다 사용되었다.

3.3 Model Structure

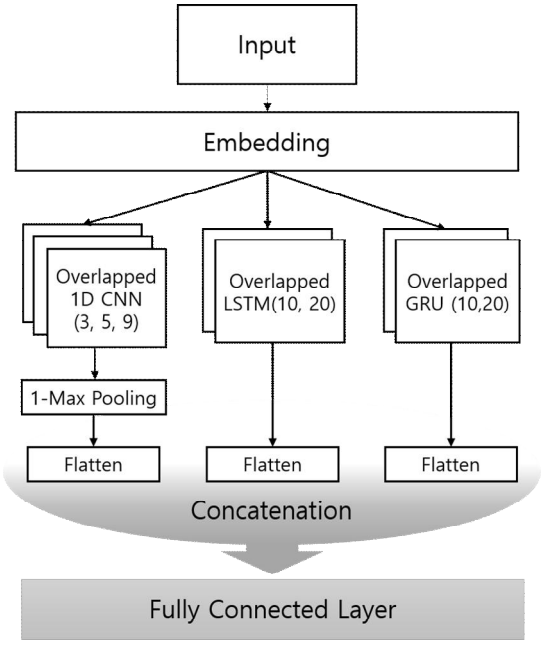

Fig. 4. Proposed model structure including overlapped RNN

Functional API을 이용해 구성된 모델의 설계는 그림 4와 같다. 동일한 입력 데이터는 1000 × 20의 형태로 임베딩 되어 모델로 들어간다. 모든 모델은 각각의 지정 수치를 가진 중첩 모델로 구성된다. 이때 출력되는 특징 맵은 알고리즘마다 형태 차원이 다르므로 다른 특징 맵과 연결하기 위해 먼저 Flatten을 수행한 뒤에 연결시켰다.

CNN 모델 부분에서는 1D CNN을 사용하며 사용된 중첩 필터의 크기는 각각 (3,5,9)이고 필터 개수는 128개이다. Filter size의 경우 비교 실험[11]에서 (3,4,5) 필터값이 가장 좋은 성능을 내었으나 짝수보다는 홀수 필터가 패턴 추출에 더 적합하다는 실험 결과와 생물학적으로 유의미한 기능을 가지는 서열 패턴의 길이는 5~7 사이인 경향이 있으므로 (3,5,9) 홀수 값으로만 재구성해서 실험하였다. 1D MaxPooling으로 피쳐의 크기를 줄인 뒤 다른 모델에서 도출된 특징 맵과 최종 합성하였다. 이 모델의 결과로 도출되는 특징 맵은 2차원이다.

RNN 계열 모델은 작은 Unit 수를 적용하고 Overlapped 시켰다. 이 중첩 모델은 LSTM과 GRU에 각각 동일하게 적용되었다. CNN 중첩 모델과 비슷한 구성을 취하되 안에 내장되어 있는 약 네 그룹 FFNN의 히든 유닛 수를 적절히 고려하여 작은 단위인 10, 20으로 설정하였다. 이 모델의 결과로 도출되는 아웃풋은 1차원 특징 맵이다. 중첩 RNN은 그림 5와 같이 도출한 벡터값을 일종의 특징으로서 중첩 CNN의 결과에 가로로 길게 합성되었다.

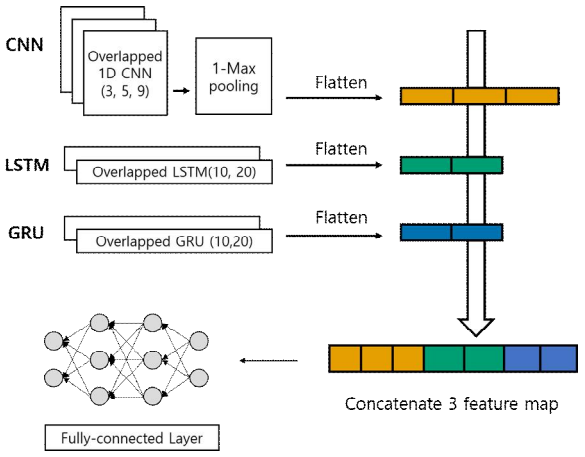

Fig. 5. LSTM/GRU output values added to the feature map

Fully-Connected Network은 비교실험에서 사용한 것과 동일하게 512개의 노드를 가지는 두 개의 층으로 구성되어있다. Activation은 ReLu이다. 최종 레이어는 Softmax로 Main Class Text Dataset과 Main Class PSSM Dataset은 총 7개 클래스를 최종 출력하며, Sub-sub Class PSSM Dataset은 139개 클래스를 최종 출력할 수 있다.

3.4 Experiment

검증 실험은 3가지 각 데이터 세트를 대상으로 수행되어졌다. 균형데이터 세트에 대해서 Loss 계산은

Categorical cross entropy가 적용되었고, 불균형 데이터 세트에 대해서는 샘플 수가 더 적은 클래스의 학습에 집중하는 Focal loss를 적용하였다. Focal loss 안에 내장된 수치는 임시 실험에 따라 알맞게 조정되었고 실험에 따라 최종적으로 감마값이 9, 알파값이 0.015로 적용되었다.

실험에 적용된 Learning rate은 0.00001과 0.000001 두 가지가 적용되었으며 처음엔 0.00001로 실험을 진행한 뒤 적용된 데이터 세트에 따라 훈련에 사용되는 데이터 개수나 훈련에 사용하는 입력 데이터 특성에 따라서 각각 다른 학습률에 최적화되었다. 최종적으로 Main Class Text Dataset에는 0.00001 학습률이, Main Class PSSM Dataset에는 0.000001 학습률이, Sub-sub Class PSSM Dataset에는 0.00001 학습률이 적용되었다.

3.5 Environment

모든 실험은 동일하게 Ubuntu 20.04, Python 3.8, RAM 256GB, NVIDIA GeForce RTX 2080 TI 4개, Tensorflow 2.4 + Keras-lr + Conda Jupter notebook 환경에서 동작하였다.

본 논문에서 설계하고 제안한 모델은 작은 학습률을 사용함에도 불구하고 과적합이 빨리 되므로 두 개의 층으로 이루어진 완전 연결 계층에서 kernel_regularizers L2를 0.00001로 설정하여 총합 2회 적용하는 것으로 과적합에 대처하였으며 완전 연결 계층 사이에 배치 정규화를 해주었고, 조기 종료를 적용하였다. 조기 종료는 검증 손실을 기준으로 판단하며 총합 5회 성능 향상이 보이지 않을 시 바로 종료하도록 설정하였다.

IV. Result and Analysis

4.1 Result

4.1.1 Model Validation with Text Dataset

표 2는 선행 연구[11]에서 가장 성능이 좋았던 Stacked LSTM과 제안 모델인 PSCREM의 Text Dataset에 대한 비교 결과이다. Text Dataset은 효소 번호 대분류까지 예측하기 위해 만들어진 데이터 세트이다. 앞으로 상술될 P Model은 제안 모델인 PSCREM을 나타내고, Lr은 학습률, ACC는 정확도, Loss는 손실률, Unit은 모델에 사용된 RNN Unit의 수를 나타낸다.

Table 2. Text Dataset Result

Model	Stacked LSTM		P Model
Lr	0.00001	0.00001	0.00001
ACC	0.8121	0.8324	0.8374
Loss	0.6974	0.5861	0.686
Unit	**20, 10**	50, 50	**20, 10**

Stacked LSTM의 형태로 20, 10 Unit을 순서대로 쌓았을 때와 이전 비교실험에서 사용되었던 Stacked LSTM 50 Unit 2회 쌓기의 결과의 손실률과 정확도를 명시하였다. 동일한 Unit을 사용했을 때 Stacked 된 LSTM의 결과보다 정확도가 2.5% 더 좋았고, 이전 비교실험에서 가장 좋았던 모델인 LSTM 50씩 2회 쌓은 모델과 비교했을 때 결과 또한 약 0.5% 더 정확도가 높았다. 그러나 손실률은 Stacked LSTM 20, 10 모델이 냈던 손실률에서 약 1% 정도 줄었고 Stacked LSTM 50 2회 모델보다는 10% 더 많았다.

4.1.2 Main class Dataset

표 3은 제안 모델의 Main Class PSSM Dataset에 대한 효소 번호 대분류 예측 3 반복 실험 결과이다. Main Class PSSM Dataset은 총 7가지 Main Class로 Label이 되어있으며 각기 1만 개씩 고르게 분포된 균형 데이터 셋이므로 Micro로 평가하지 않고 Macro precision, recall, F1-Score로 평가하였다.

Table 3. Main class PSSM Dataset Result

	P Model		
# of experiment	1	2	3
Loss	0.5121	0.5038	0.5007
ACC	0.8621	0.8641	0.8662
Macro_precision	0.8629	0.8648	0.8663
Macro_recall	0.8624	0.8644	0.8652
Macro_F1	0.8625	0.8646	0.8655

표 3에서 # of experiment는 실험의 회차를, ACC는 정확도를 나타낸다. 해당 실험에서 학습률은 0.000001로 적용하고 kernel_regularizers 는 L2 0.00001을 적용했을 때 가장 학습 진행이 좋았다.

실험을 3회 반복한 뒤 그 평균값을 취하면 정확도는 86.4% ± 0.2이고, 손실률은 50.5% ± 0.7이었다. Main class Text Dataset 제안 모델 결과에 비해 약 2.7% 정도 더 정확도가 높았고 Loss 값은 약 18% 정도 더 낮았다. Text Dataset의 Stacked LSTM 50, 50 결과와 비교하면

약 3.2% 정도 정확도가 더 높았고, 손실률은 약 8% 더 낮았다.

3번의 실험 모두 조기 종료 함수에 의해 조기 종료되었으며 100 epoch를 적용하였음에도 첫 번째 실험은 30 epoch에서, 두 번째 실험은 29 epoch에서, 세 번째 실험은 30 epoch에서 조기 종료되었다.

4.1.3 Sub-Sub class Dataset

표 4는 세분류 예측 문제에 대한 제안 모델의 결과를 나타낸 것이다. # of experiment는 실험의 번수를, ACC는 정확도를, Loss는 손실률을 나타낸다.

Table 4. Sub-Sub Class Dataset Result

	P Model		
# of experiment	1	2	3
Loss	0.3138	0.3045	0.3258
ACC	0.8452	0.8486	0.8387
Micro_precision	0.84523	0.84857	0.83873
Micro_recall	0.84523	0.84857	0.83873
Micro_F1	0.84523	0.84857	0.83873

Sub-Sub Class PSSM Dataset는 각 1만 개였던 Main Class Dataset과 달리 각 데이터 세트를 50%씩 랜덤하게 재샘플링하여 매 실험마다 총개수 Train 70,682, Validation 23,646, Test 23,646개씩 사용되었다. 충분한 학습 데이터가 존재하므로 해당 실험에서 학습률은 0.00001로 앞선 대분류 예측 문제보다는 조금 높여서 적용하였고 0.000001을 적용하였을 때보다 0.00001을 적용했을 때 학습 진행이 가장 좋았다. 클래스 간의 극심한 불균형을 고려하여 평가척도는 Micro로 정하였다.

표 4에서 확인됨과 같이 3회 반복된 세분류 예측 실험에서 제안 모델의 평균 정확도는 84.4% ± 0.4이고, 평균 손실률은 31.5% ± 1이었다. 3번의 실험 모두 조기 종료 함수에 의해 조기 종료되었으며 50 epoch를 적용하였음에도 첫 번째 실험은 10 epoch에서, 두 번째 실험은 10 epoch에서, 세 번째 실험은 11 epoch에서 조기 종료되었다. 각 클래스 당 7만 개인 대분류 예측 문제보다 학습에 참고할 데이터의 양이 많아 조금 높은 학습률을 적용하였으나 0.000001을 사용하였을 때보다 0.00001 학습률을 사용하여 학습했을 때 모델의 학습 결과나 진행 정도가 가장 완만하고 좋았다.

4.1.4 Comparison with Benchmarks

대상 도구는 다음과 같다. 원-핫 인코딩과 Overlapped CNN만을 사용한 DeepEC, 효소분류 7번을 고려한 MF-EFP, PSSM을 사용한 기존 선행 연구인 ECPred로 총 3가지이다. 비교 상의 모델 중 MF-EFP만이 2018년 8월 이후 EC 7이 새 대분류로 분화된 것을 고려하여 만들어진 효소 번호 예측 모델이다. 공정한 평가를 위해 동일한 데이터로 결과를 비교하였다. 비교를 위해 사용한 데이터는 효소 번호 대분류까지 예측하는 7 분류 문제이며, 90번 대 효소 번호가 포함되지 않고, 중복된 효소 번호를 가지지 않는 237,923개의 서열에서 각 클래스에 해당하는 서열을 200개씩 무작위로 추출하였다.

Table 5. Benchmark results for Dataset including EC 7

	DeepEC	MF-EFP	ECPred	P Model
ACC	73.64%	17.4%	82.1%	89.2%
Macro F1	0.6473	0.2029	0.6047	0.8914

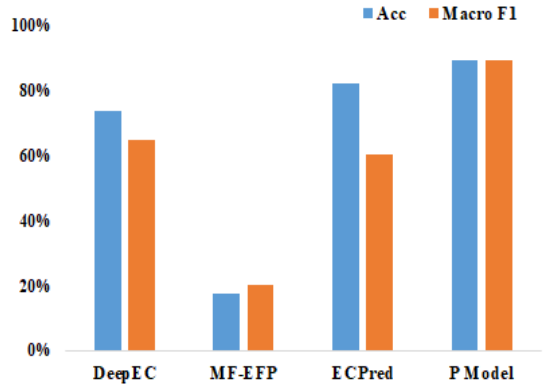

Fig. 6. Comparison 4 Tools Graph for Dataset including EC 7

표 5는 EC 7번이 포함된 데이터를 사용하여 EC 번호 대분류까지 각 모델이 예측한 결과이다. 그림 5는 표 5를 시각화한 자료이다. 정확도와 Macro F1 점수를 사용하여 그 성능을 평가하였다. DeepEC가 73.64%, MF-EFP가 17.4%, ECPred가 82.1%, 제안 모델이 89.2%로 제안 모델의 정확도가 가장 높았다. 7번은 전혀 예측하지 못하는 다른 모델에 비해 효소 번호 7번까지 고려한 MF-EFP와의 정확도 차이가 72%로 가장 컸다. Macro F1으로 분류 성능을 나타내었을 때 DeepEC가 64%, MF-EFP가 20%, ECPred가 60%, 제안 모델이 89%로 이 역시 제안 모델의 매크로 F1 점수가 다른 2개의 Tool에 비해 더 높았다.

Table 6. Benchmark results for Dataset not including EC 7

	DeepEC	MF-EFP	ECPred	P Model
ACC	85.91%	20.1%	95.8%	87.8%
Macro F1	0.7840	0.2290	0.7277	0.7568

표 6은 EC 7번을 제거한 데이터를 사용하여 EC 번호 대분류까지 각 모델이 예측한 결과이다. DeepEC와 ECPred가 7번을 학습에 포함하지 않았기에 MF-EFP와 제안 모델 또한 7번을 포함하지 않은 데이터 셋으로 비교 실험을 한 번 더 수행하였다. 해당 실험 결과 또한 정확도와 Macro F1 점수를 사용하여 평가하였다. 그림 6으로 시각화하였다.

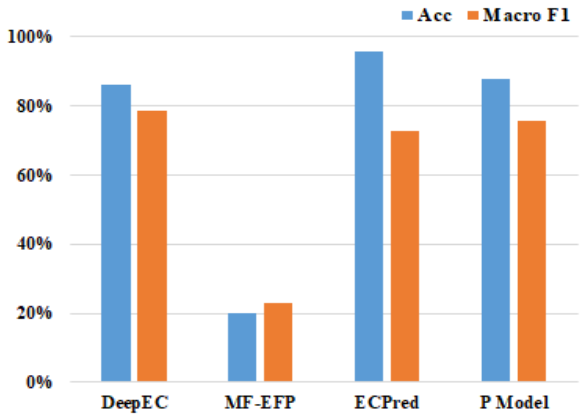

Fig. 7. Comparison 4 Tools Graph for Dataset not including EC 7

DeepEC가 85.91%, MF-EFP가 20.1%, ECPred가 95.8%, 제안 모델이 87.8%로 효소 7번이 제외된 데이터를 대상으로는 ECPred의 정확도가 가장 높았고 제안 모델의 정확도가 두 번째로 높았다. 그러나 Macro F1으로 나타내었을 때 ECPred가 72.7%로 순수정확도에 비하여 낮은 수치의 값을 보여주었고, 75.6%인 제안 모델의 결과보다 낮아졌으며 DeepEC가 78.4%, MF-EFP가 22.9%로, 정확도만을 따졌을 때 제안 모델보다 약 2% 낮았던 DeepEC의 값이 4가지 도구 중에서 가장 Macro F1 점수가 좋았다. 이 역시도 제안 모델이 두 번째로 좋았다.

표 5와 표 6을 전부 포함하여 정확도와 매크로 F1 점수로 평균 순위를 취하였을 때 DeepEC가 2.3, MF-EFP가 4, ECPred가 2, 제안 모델이 1.06으로 제안 모델의 평균 순위가 가장 높았고, 표 6만으로 정확도와 매크로 F1 점수의 평균 순위를 취하였을 때 MF-EFP의 순위가 4, DeepEC, ECPred, 제안 모델의 순위가 2로 모두 같았다.

4.2 Discussion

본 연구에서 제안한 모델의 성능 실험 결과 효소 번호 첫 번째 자리를 예측하는 문제에 대해서는 평균 86.4%의 정확도를 나타냈고, 오차범위는 ±0.2% 이내였다. 효소 번호 3번째 자리까지 예측하는 문제에 대해서는 평균 정확도 84.4%였으며 오차범위는 ±0.6% 이내였다.

4.1.1의 첫 번째 Text Dataset을 이용한 검증 실험에서 이전 비교실험에서 가장 좋았던 모델인 LSTM 50씩 2회 쌓은 모델과 제안 모델의 결과를 비교했을 때 약 0.5% 더 정확도가 높았고 동일한 유닛 수로 Stacked 된 LSTM의 결과보다는 2.5% 더 좋았다. 비록 손실률은 Stacked LSTM 20, 10 모델이 냈던 손실률에서 약 1% 정도 줄어들었던 것에 비해 Stacked LSTM 50 2회 모델보다는 10% 더 많았으나 해당 실험에서 Unit을 10, 20으로 적용했을 때 일반적으로 사용되는 50 unit 2회 쌓기와 비슷한 결과를 도출해내었으므로 이는 Unit을 50, 70으로 적용했을 때 100 unit 2회 쌓기와 비슷한 결과를 도출하거나 더 좋은 성능을 보여줄 것으로 보였다.

4.1.2에서 보인 Main Class PSSM Dataset을 이용한 두 번째 검증 실험에서는 Text Main Class Dataset을 이용한 첫 번째 실험보다 평균 정확도가 약 2.7% 더 좋았고 손실률은 18% 낮았다. Stacked 50, 50 결과와 비교하면 약 3.2% 정도 정확도가 더 높았고, 손실률은 약 8% 더 낮았다. 첫 번째 데이터 세트인 Main class Text Dataset에는 원-핫 임베딩이 적용되었고 나머지 Main Class, Sub-Sub Class PSSM Dataset에는 PSSM이 사용되었다. 두 실험에서 사용한 제안 모델의 구성 및 데이터는 전부 동일하다. 이를 통해 모델에 서열 데이터를 입력할 때 일반 원-핫 임베딩보다 PSSM을 통해 서열 진화정보를 사용하였을 때 예측 모델의 성능이 약 3% 더 올라감을 확인하였다. 이는 PSSM의 내부 행렬 안에 보호된 서열의 진화정보가 서열 내의 구조에 관련된 중요한 패턴을 포함하는 것처럼 기능에 관한 중요한 패턴도 반드시 정보로서 요약하기 때문이다. 이는 효소의 구조가 기능과 밀접한 연관성을 지니는 고유 특성과도 밀접하게 관련된다.

4.1.3을 통해 확인할 수 있는 Sub-Sub Class PSSM Dataset를 이용한 검증 실험에서는 데이터가 비록 불균형하나 Focal Loss로 수가 더 적은 클래스에 집중해서 가중치를 높이는 방향으로 학습시킴으로서 불균형 데이터 클래스의 편향을 일부 해소하였다. Focal loss를 적용하지 않았을 때의 세분류 예측 문제의 손실률과 정확도는 항상 1 이상, 50% 미만이었다. 세분류 예측 문제에 대한 제안 모델의 손실률과 정확도는 각각 31.5%, 84.4%였으며 클

래스가 132개 더 증가하였으나 정확도가 7 분류 문제인 Main class dataset과 비슷한 성능을 보였다. 또한 마지막 검증을 위해 DeepEC, MF-EFP, ECPred를 직접 실험하여 제안 모델과 비교하였다. PSCREM에서 처음으로 사용된 Overlapped RNN이라는 실험적인 특징 맵 합성법에도 불구하고, 제안 모델은 다른 연구가 공표한 단백질 기능 예측 실전 모델과 실 예측 성능이 크게 차이 나지 않음을 확인하였다. 특히 4.1.4의 결과를 보면 PSCREM과 가장 유사한 조건을 가진 툴과의 성능 차이는 확연하였다. 효소 번호 대분류 예측 실험 결과에서 대분류 7번까지 고려한 MF-EFP와 비교하면 72% 이상 크게 향상된 성능을 보여주었다. 성능에 관해 정확도와 F1 점수로 평균 순위를 취하였을 때 제안 모델은 2순위로 항상 일정하였으며 분류 개수에 상관없이 약 85% 정도 늘 일정한 성능을 보였다.

일반적으로 모델이 분류해야 할 Class가 적으면 해당 데이터에 과적합 되고, 분류해야 할 Class가 늘면 성능이 다소 떨어지는 경향을 보이나 본 논문에서 제안한 모델은 그렇지 않았다. 이는 제안 모델이 서열 진화정보를 입력으로 사용함으로써 서열 간 아미노산의 위치가 우연히 일치할 수 있는 경우의 수인 확률적 노이즈 값에 대한 영향을 덜 받기 때문이며, RNN 계열의 문맥 정보를 특징 맵의 일부로 사용함으로써 학습에 참조할 정보가 늘어 기능이 다른 단백질 서열이 가지는 고유 패턴 차이를 더 잘 구분 지을 수 있게 되었다는 것을 의미한다. 이로 인해 모델은 클래스가 늘더라도 늘 일정한 성능을 가질 수 있게 된다.

V. Conclusions

본 모델에서 처음으로 시도된 Overlapped RNN은 단백질 기능 및 구조 예측 추출 분야에 새로운 방법론으로서 제안된다. LSTM과 GRU가 각기 다르게 파악한 문장 간 관계 값을 통합 특징 맵에 더함으로써 모델의 패턴 학습시 참조할 정보를 보충하였고, 작은 unit을 가지고도 약 2배 이상 되는 Unit 수를 가진 Stacked RNN 계열과 비슷한 성능을 낼 수 있음을 실험으로 확인하였다.

또한 본 연구는 딥러닝을 사용한 단백질 기능 예측 문제에 서열 진화정보 사용의 중요성을 입증하였다. 기존의 원-핫 임베딩은 쉽고 빠르나 단백질 서열 내의 기능과 관련된 세부 정보가 CNN의 합성곱 과정에서 다수 누락될 수 있다는 단점이 있었다. 그러나 제안 모델에서는 서열의 세부 정보를 보존하는 PSSM을 사용함으로써 모델은 서열상 기능과 관련된 세세한 특이 맥락을 보호하고 기능적 세부 패턴을 유지하였다. 그 결과 모델의 성능은 향상되었다. 그러나 PSSM 제작에 소요되는 시간이 길어 PSSM profile 제작 시간 단축에 관한 연구가 필요하다.

본 제안 모델에 입력으로 사용된 PSSM profile은 기능보다는 구조 예측 연구에 더 자주 사용되곤 하는 profile 방법이나 구조와 기능이 아주 밀접하게 연관된 효소의 특성에 기반을 두어 실험과 제안 모델을 설계하였으며, 실험을 통해 서열 진화정보를 이용한 효소 기능 예측이 가능할 뿐만 아니라 충분히 유용하다는 것을 증빙하였다.

ACKNOWLEDGEMENT

* This research was supported by the MSIT(Ministry of Science, ICT), Korea, under the High-Potential Individuals Global Training Program)(2021-0-01581) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

* This research was supported by the BK21 FOUR (Fostering Outstanding Universities for Research) funded by the Ministry of Education(MOE, Korea) and National Research Foundation of Korea(NRF)

REFERENCES

[1] Y. Liang, S. Liu, S. Zhang, "Prediction of Protein Structural Classes for Low-Similarity Sequences Based on Consensus Sequence and Segmented PSSM", Computational and Mathematical Methods in Medicine, vol. 2015, 9 pages, Dec, 2015. https://doi.org/10.1155/2015/370756

[2] J. Wang, B. Yang, J. Revote, A. Leier, T. T Marquez-Lago, G. Webb, J. Song, K. Chou, T. Lithgow, “POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles”, Bioinformatics, Volume 33, Issue 17, 01 September 2017, Pages 2756–2758, https://doi.org/10.1093/bioinformatics/btx302

[3] Mousavian Z, Khakabimamaghani S, Kavousi K, Masoudi-Nejad A., “Drug-target interaction prediction from PSSM based evolutionary information.”, Journal of pharmacological and toxicological methods, vol. 78, 42-51, March–April, 2016, doi:10.1016/j.vascn.2015.11.002

[4] N. Q. K. Le and V. N. Nguyen. "SNARE-CNN: a 2D convolutional neural network architecture to identify SNARE proteins from

high-throughput sequencing data." PeerJ. Computer science, vol. 5, e177, Feb, 2019, doi:10.7717/peerj-cs.177

[5] Y. Guo, J. Wu, H. Ma, S. Wang, and J. Huang, “EPTool: A New Enhancing PSSM Tool for Protein Secondary Structure Prediction“, Journal of computational biology : a journal of computational molecular cell biology, vol. 28, 362-364, Apr, 2021, doi:10.1089/cmb.2020.0417

[6] Liu Y, Gong W, Yang Z, Li C., “SNB-PSSM: A spatial neighbor-based PSSM used for protein-RNA binding site prediction.”, J Mol Recognit, vol.34, e2887, June, 2021, https://doi.org/10.1002/jmr.2887

[7] A. Dalkiran, A. S. Rifaioglu and M. J. Martin et al, “ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature.”, BMC bioinformatics, vol. 19, 334, Sep, 2018, https://doi.org/10.1186/s12859-018-2368-y

[8] A. Amidi, S. Amidi and D. Vlachakis et al, “EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation.”, PeerJ, vol. 6, e4750, May, 2018, doi:10.7717/peerj.4750

[9] X. Xiao, L. Duan and G. Xue et al, "MF-EFP: Predicting Multi-Functional Enzymes Function Using Improved Hybrid Multi-Label Classifier", in IEEE Access, vol. 8, pp. 50276-50284, Mar, 2020, 10.1109/ACCESS.2020.2979888

[10] N. Strodthoff, P. Wagner, M. Wenzel and W. Samek, “UDSMProt: universal deep sequence models for protein classification”, Bioinformatics, Vol 36(8), 2401–2409, Apr, 2020, https://doi.org/10.1093/bioinformatics/btaa003

[11] J. Lee, H. Lee, “Comparison of Deep Learning Models Using Protein Sequence Data”, KIPS Transactions on Software and Data Engineering, Vol. 11, No. 6, pp. 245-254, Jun, 2022, https://doi.org/10.3745/KTSDE.2022.11.6.245

[12] Suzuki H (2015). "Chapter 7: Active Site Structure". How Enzymes Work: From Structure to Function. Boca Raton, FL: CRC Press. pp. 117–140. ISBN 978-981-4463-92-8.

[13] D. M. Debra, “Enzyme function discovery.”, Structure, vol. 16(11), 1599-600, NOV, 2008, doi:10.1016/j.str.2008.10.001

[14] Saigo, Hiroto et al. “Reaction graph kernels predict EC numbers of unknown enzymatic reactions in plant secondary metabolism.”, BMC Bioinformatics, 11 Suppl 1(Suppl 1), S31, Jan, 2010, doi: 10.1186/1471-2105-11-S1-S31.

[15] A. G. McDonald and K. F. Tipton, “Enzyme nomenclature and classification: the state of the art.”, FEBS J, Nov, 2021, doi.org/10.1111/febs.16274

[16] A. A. Schäffer 1, L. Aravind, T. L. Madden, " Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements.", Nucleic Acids Res, vol. 29(14), 2994-3005, Jul, 2001, doi: 10.1093/nar/29.1.2994.

[17] S. Kim, "Basic for Protein Structure Prediction: BLAST and Profile", Biophysical Society Newsletter, vol. 11, no. 1, October 2005.

[18] Y. Kim, "Convolutional Neural Networks for Sentence Classification", In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746-1751, Oct, 2014, 10.3115/v1/D14-1181

[19] J. Y. Ryu, H. U. Kim, S. Y. Lee, "Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers", Proceedings of the National Academy of Sciences of the United States of America, 116 (28), 13996-14001, June, 2019, https://doi.org/10.1073/pnas.1821905116

[20] Gao, Ruibo et al. “Prediction of Enzyme Function Based on Three Parallel Deep CNN and Amino Acid Mutation.” International journal of molecular sciences, vol. 20(11), 2845, Jun, 2019, doi:10.3390/ijms20112845

[21] A. L. Rio, M. Martin, A. Perera-Lluna and R. Saidi , "Effect of sequence padding on the performance of deep learning models in archaeal protein functional prediction.", Scientific Reports, 10(1), 14634, Sep, 2020, https://doi.org/10.1038/s41598-020-71450-8

Authors

Jeung Min Lee received the B.S degrees in Department of Computer Science and Engineering from Sunmoon University, Korea, in 2017. She received the M.S. degrees in Bio Big Data Convergence Major, Department of Computer and Electronics Convergence Engineering, Sunmoon University, Korea, in 2022. Jeung Min Lee is interested in Deep Learning in Bioinformatics, statistical inference and models for diverse networks in Biomedical sciences.

Hyun Lee received the B.S degrees in Division of Computer Science and Engineering from Sunmoon University, Korea, in 2000. He received the M.S. degrees in Department of Management Computer Science and Engineering from Sunmoon University, Korea in 2002. He received the Ph.D. Computer Science and Engineering from the Univ. of Texas at Arlington, Arlington, U.S. in 2010. Dr. Lee is currently a Professor in the Department of Computer Science and Engineering from Sunmoon University. He is interested in Decision Support System, Autonomic Computing, Elder Human Care System, and Internet of Things-based Cyber physics (CPS) system.