Bi-SeqCNN: A Novel Light-weight Bi-directional CNN Architecture for Protein Function Prediction : Supplemental document S1

Vikash Kumar, Akshay Deepak, Ashish Ranjan, and Aravind Prakash

In this supplement file, we have added three sections (Sections S.I to S.III). Training vs. validation loss curve in Section S.I on CAFA3 and Data2016 datasets, choosing different CNN parameters in Section S.II, and lastly, performance analysis of the proposed ensemble model on varying embedding dimensions in Section S.III.

S.I. Training vs. Validation Loss Curve

In this section, we cover the training and validation loss characteristics with epochs for the proposed model Bi-SeqCNN on the CAFA3 and Data2016 datasets (explained in the main manuscript, in Section 3.1), as shown in Fig. S1. All of the graphs in Fig. (S1a to S1f) are considered good fits since almost all of them show comparable behavior, with both the training and validation losses convergent at a specific point.

S.II. Choosing the CNN parameters

In this section, we discuss the parameters selection with the proposed *Bi-directional CNN*-based method (shown in the main manuscript, in Fig. 3) on benchmark protein sequence datasets (discussed in the main manuscript, in Section 3.1). This include varying the CNN parameters, such as, filter size and the number of filters.

A. Results based on varying the filter-size

Here, the filter-size is varied $\in$ {4, 5, 6 } while fixing the embedding dimension = 32, number of filters = 64, and segment size = 150. The experimental results are shown in Table I. There is no single filter size that produces superior results. However, overall filter size = 6 produces more number of superior results in Table I.

B. Results based on varying the number of filters

Table II shows the experimental results of the proposed model varying the number of filters as 64, 128 and 256, while fixing the filter-size = 6, embedding dimension = 32, and segment size= 150. There is no single value for the number of filters that produces superior results on all three datasets and all three sub-ontologies. However, number of filters = 256 overall produces more number of superior results.

C. Section Summary

Overall, with filter size = 6 and # Filters = 256, the proposed method produces superior results as shown in Tables I and II. Hence, proposed model with these configurations are used for the experiments in the remaining part of the paper.

S.III. Performance analysis of the proposed ensemble model on varying embedding dimension

In this section, we discuss about the results produced by the proposed ensemble model $Bi\text{-}SeqCNN^{+}$ (as discussed in the main manuscript, in Section 2.3) on varying the embedding dimension $\in$ {32, 48, and 64}. The results are shown in Table III.

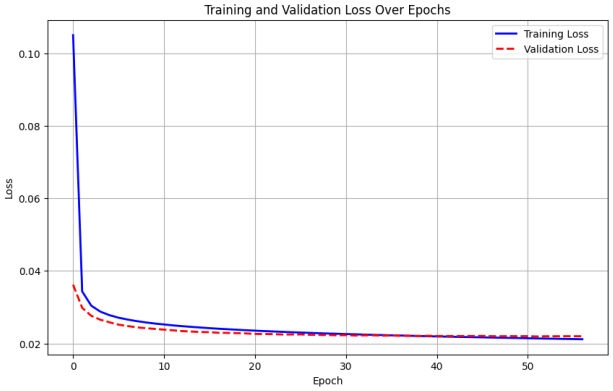
The F_{max} progressively increases on increasing the embedding dimension for the $Bi\text{-}SeqCNN^{+}$ model on Data2017 [*S1*] for both BP and MF sub-ontologies. It produces best F_{max} score with 64 embedding dimension. Further, on CAFA3 dataset [*S2*], highest F_{max} score is obtained with 64 (for BP and CC sub-ontologies) and 48 (for MF sub-ontology) embedding dimension. Lastly, for Data2016 [*S2*], $Bi\text{-}SeqCNN^{+}$ produces the best F_{max} score on BP and MF sub-ontologies dataset with 64 embedding dimension.

Overall, in six out of eight cases, embedding dimension = 64 shows the best performance for the proposed ensemble model, and hence, is used in sections to follow.

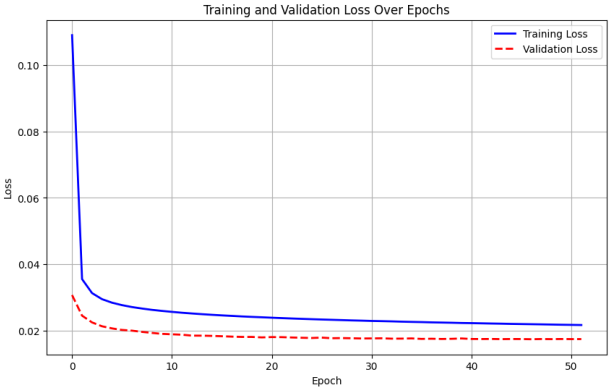
References

[S1] Ranjan, A., Tiwari, A. and Deepak, A., (2021). "A sub-sequence based approach to protein function prediction via multi-attention based multi-aspect network". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, doi: 10.1109/TCBB.2021.3130923. Epub 2021 Nov 26. PMID : 34826296.

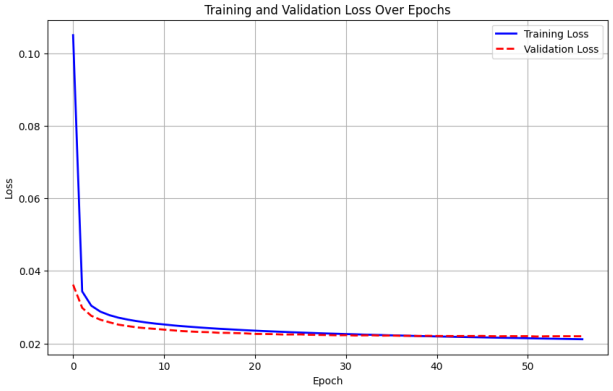
[S2] Kulmanov, M. and Hoehndorf, R., (2020). "DeepGOPlus: improved protein function prediction from sequence". *Bioinformatics*, 36(2), pp.422-429.



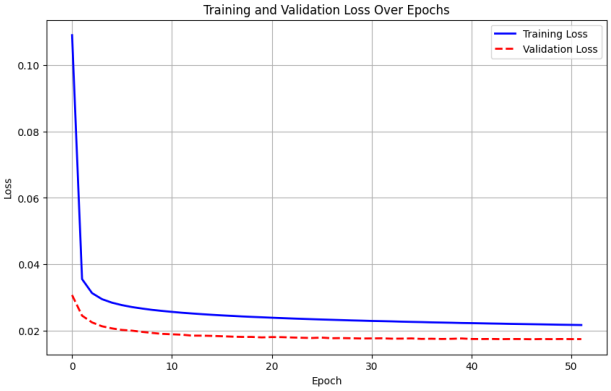
(a) CAFA3-BP Data



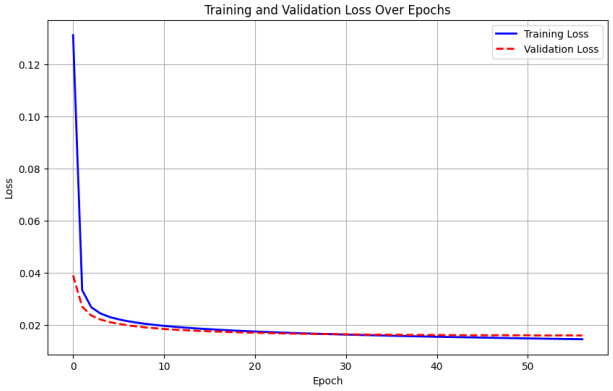
(b) Data2016-BP Data



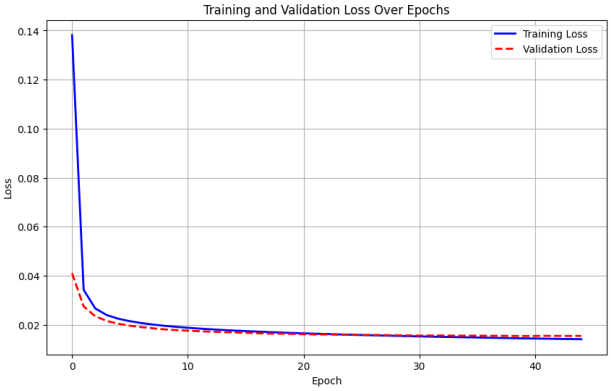
(c) CAFA3-CC Data



(d) Data2016-CC Data



(e) CAFA3-MF Data



(f) Data2016-MF Data

Fig. S1. Training and validation losses over epochs curve for CAFA3 and Data2016 datasets using the proposed Bi-SeqCNN model. The Data2016 dataset's training and validation loss over epochs curve is plotted on the right, while the CAFA3 dataset's training and validation loss over epochs curve is displayed on the left.

TABLE I
COMPARISON ANALYSIS OF DATA2017, CAFA3, AND DATA2016 DATASETS FOR DIFFERENT SUB-ONTOLOGY CLASSES AGAINST Pre_{Avg}, F_{max}, AND AUPR PERFORMANCE METRICS WITH #FILTERS = 64, EMBEDDING DIMENSION = 32, AND SEGMENT SIZE = 150.

Datasets —>			Data2017			CAFA3			Data2016		
S. No.	Sub-ontology	Filter-size	Pre_{Avg}	F_{max}	AUPR	Pre_{Avg}	F_{max}	AUPR	Pre_{Avg}	F_{max}	AUPR
1	BP	4	0.454	0.451	0.248	0.532	0.484	0.472	0.445	0.411	0.386
		5	0.429	0.446	**0.251**	0.530	0.486	0.467	**0.461**	0.407	0.382
		6	**0.460**	**0.453**	0.248	**0.541**	**0.490**	**0.477**	0.437	**0.411**	**0.389**
2	CC	4	-	-	-	0.653	0.626	**0.642**	0.696	**0.682**	**0.703**
		5	-	-	-	**0.681**	**0.630**	0.638	**0.716**	0.682	0.701
		6	-	-	-	0.674	0.628	0.639	0.693	0.680	0.701
3	MF	4	0.566	0.581	0.379	**0.615**	0.488	**0.457**	0.628	0.554	0.514
		5	**0.576**	**0.591**	0.390	0.580	0.492	0.456	0.611	0.550	0.508
		6	0.568	0.588	**0.403**	0.579	**0.493**	0.455	**0.659**	**0.556**	**0.519**

TABLE II
COMPARISON ANALYSIS OF DATA2017, CAFA3, AND DATA2016 DATASETS FOR DIFFERENT SUB-ONTOLOGY CLASSES AGAINST Pre_{Avg}, F_{max}, AND AUPR PERFORMANCE METRICS WITH FILTER-SIZE = 6, EMBEDDING DIMENSION = 32, AND SEGMENT SIZE = 150.

Datasets —>			Data2017			CAFA3			Data2016		
S. No.	Sub-ontology	#Filters	Pre_{Avg}	F_{max}	AUPR	Pre_{Avg}	F_{max}	AUPR	Pre_{Avg}	F_{max}	AUPR
1	BP	64	**0.456**	0.453	0.248	0.541	0.490	0.477	0.437	0.411	0.389
		128	0.449	0.453	0.276	0.549	0.495	0.478	**0.465**	0.415	0.389
		256	0.443	**0.457**	**0.281**	**0.557**	**0.499**	**0.482**	0.450	**0.418**	**0.394**
2	CC	64	-	-	-	**0.674**	**0.628**	**0.639**	0.693	0.680	0.701
		128	-	-	-	0.672	0.627	0.637	**0.716**	0.683	0.701
		256	-	-	-	0.661	0.627	0.635	0.702	**0.690**	**0.704**
3	MF	64	**0.568**	**0.588**	0.403	0.579	0.493	0.455	0.659	0.556	0.519
		128	0.566	0.582	0.374	**0.600**	0.499	0.468	**0.673**	**0.567**	0.521
		256	0.561	0.581	**0.484**	0.584	**0.506**	**0.468**	0.641	0.565	**0.532**

TABLE III
CLASSIFICATION RESULTS OF THE PROPOSED ENSEMBLE MODEL WITH VARYING EMBEDDING DIMENSION (DIMENSION: EMBEDDING DIMENSION).

Dataset —>			Data2017			CAFA3			Data2016		
S. No.	Sub-Ontology	Dimension	Pre_{Avg}	F_{max}	AUPR	Pre_{Avg}	F_{max}	AUPR	Pre_{Avg}	F_{max}	AUPR
1	BP	32	0.674	0.548	0.497	0.552	0.514	0.505	0.473	0.429	0.412
		48	0.711	0.573	0.511	0.559	0.523	0.514	**0.495**	0.434	0.416
		64	**0.734**	**0.583**	**0.525**	**0.593**	**0.524**	**0.517**	0.474	**0.441**	**0.423**
2	CC	32	–	–	–	0.654	0.637	0.674	**0.721**	0.705	0.747
		48	–	–	–	**0.683**	0.640	**0.675**	0.711	**0.707**	**0.748**
		64	–	–	–	0.645	**0.642**	0.674	0.719	0.706	**0.748**
3	MF	32	**0.827**	0.675	0.656	0.607	0.533	0.532	**0.698**	0.591	0.581
		48	0.815	0.692	0.652	**0.649**	**0.537**	**0.534**	0.670	0.586	0.587
		64	0.826	**0.696**	**0.662**	0.614	0.535	**0.534**	0.679	**0.593**	**0.589**