Convolutional Neural Network Architectures for CAFA4

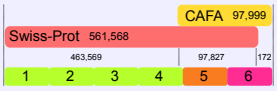
Jari Björne
Department of Future Technologies,
University of Turku, FI-20014, Finland
jari.bjorne@utu.fi

Introduction

The Critical Assessment of protein Function Annotation (CAFA) 4 challenge concerns the prediction of annotated ontology terms for given protein sequences, using the Gene Ontology (GO), the Human Phenotype Ontology (HPO) and the Disorder Ontology (DO), with 100k sequences provided as prediction targets [1].

In this work convolutional neural network architectures are applied for the task [2]. UniProt sequence data is enriched with taxonomical information and InterProScan analyses before being processed with neural networks, predicting the different ontologies' terms as a multi-label classification task [3,4].

Dataset



The dataset consists of the 560k union of UniProt and CAFA4 prediction targets. GO and lineage annotations are imported from the UniProt XML release and HPO and DO annotations from their respective databases.

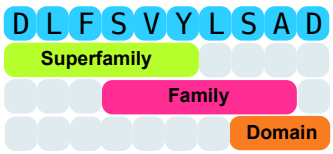
For training, the dataset is divided into six subsets, with all homologs of a protein in the same subset. In the six-fold cross-validation four subsets at a time are used for training, one for parameter optimization and one for prediction.

Evaluation

The 1000 most common terms are predicted in a multilabel approach. Performance is evaluated using the F1 and AUC scores. The six-fold cross-validation predicts the entire dataset, with CAFA4 targets separated for submission.

For labels, either all annotations are used or only ones not marked as IEA (inferred from electronic annotation). Experiments are also performed with either all sequences or only sequences with at least one annotated term (with the goal of reducing false negatives).

System Overview

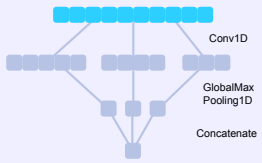


A maximum of 1000 amino acids are used per sequence. Embeddings are genered for these and the InteProScan sequence features. The sequence mapped into sets of embeddings and processed with a dropout layer forms the input of the convolutional network. Four different convolutional network architectures are tested for term prediction. The system is implemented using the Keras package [5].

Dropout 0.1

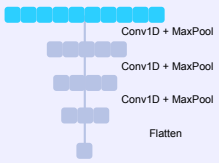
Parallel CNN

A set of 1D convolutions are run on the input and max pooled together.



Nested CNN

A set of 1D convolutions followed by max pooling are run one after the other.



MobileNet V2

The MobileNet architecture is a fast and performant solution for image analysis [6]. A 1D version of this network architecture is tested for protein function prediction.

DeepGOPlus

A state-of-the-art protein function prediction network, based on multiple parallel convolutions with a large filter set [7].

Taxonomic lineage features are inserted into the model after the main convolutional network.

Eukaryota
Vertebrata
Primates
Homo

Dropout 0.1

Dense 1000

Labels 1000

Highest Performing Terms

P	R	F	A	TP	FP	TN	FN	GO	TERM
0.700	0.831	0.760	0.915	2204	945	558144	447	0004674	protein serine/threonine kinase activity
0.712	0.814	0.760	0.906	2801	1133	557164	642	0004672	protein kinase activity
0.654	0.594	0.623	0.796	2124	1124	557041	1451	0006468	protein phosphorylation
0.696	0.560	0.621	0.779	2800	1223	555517	2200	0016773	phosphotransferase activity, alcohol group as acceptor
0.694	0.487	0.572	0.742	2806	1238	554740	2956	0016301	kinase activity
0.794	0.443	0.569	0.721	755	196	559842	947	0004930	G protein-coupled receptor activity

Sample terms are shown for the submitted model TFP3, based on the DeepGoPlus architecture. All annotations and all sequences were used in the training of this model.

P = Precision R = Recall
F = F-score A = AUC
TP/FP = True/False Positive
TN/FN = True/False Negative

References

[1] **A large-scale evaluation of computational protein function prediction.** Radivojac, Predrag et al. *Nature methods (2013) 10 (3), pp.221-227.*
[2] **Y. LeCun, Y. Bengio, et al.** Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks, 3361(10):1995, 1995.*
[3] **UniProt Consortium.** Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research, 47 (D1):D506–D515, 2019.*
[4] **P. Jones, D. Binns, et al.** InterProScan 5: genome-scale protein function classification. *Bioinformatics, 30(9): 1236–1240, 2014.*
[5] **Keras.** Chollet, François et al. *https://github.com/fchollet/keras (2015).*
[6] **M. Sandler, A. Howard, et al.** MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE CVPR, pp. 4510–4520, 2018.*
[7] **M. Kulmanov and R. Hoehndorf.** DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics, 36(2):422–429, 2020.*

Turun yliopisto
University of Turku