DOI 10.18699/SBB-2022-668

Functional domain annotation of protein sequences with deep metric learning

Russkikh N.[1, 2, 3]*, Ryazantsev G.[2], Antonets D.[1, 2, 3], Shaburova E.[1, 2], Vyatkin Yu.[2, 4]

[1] *State Research Center of Virology and Biotechnology "Vector", Koltsovo, Novosibirsk, Russia*

[2] *AcademGene LLC, Novosibirsk, Russia*

[3] *A.P. Ershov Institute of Informatics Systems, SB RAS, Novosibirsk, Russia*

[4] *Novosibirsk State University, Novosibirsk, Russia*

* *russkikh@nprog.ru*

Key words: representation learning, protein functional annotation, domain, deep metric learning

Motivation and Aim: Functional domain annotation of amino acid sequences is a very important and long-standing task. Identification of functional domains within the protein can provide us insights not only into its functions but also into its origin and evolution. Current methods of choice, such as PHMMER [1], HHSEARCH [2] etc. are based on homology detection using hidden Markov models (HMMs). Despite the achieved progress, the task remains challenging in certain cases, especially when dealing with sequences with very little or no homology towards the annotated proteins. We propose a deep metric learning approach to address this issue. Benchmark on remote homology task constructed from Pfam-seed dataset demonstrated that the proposed method makes fewer errors than the conventional methods and performs on par with deep learning-based methods while being able to predict multiple domains in the protein at once.

Methods and Algorithms: We treat protein functional annotation problem as a tokenwise classification deep learning task. Our model consists of a multilayer transformer encoder with a small two-layer perceptron on top of its output. As a transformer backbone, we used ProtBert [3]. Output features from four ProtBert intermediate layers, from 10 to 13, are concatenated and passed to the feedforward head, consisting of two layers with 1024 and 64 neurons respectively, with ReLU activation in between them. The model was trained on Pfam-seed dataset using soft triplet loss [4]. In order to classify the positional tokens, we used the NED (Normalized sum of Exponential of the Distances) algorithm [5].

Results: In this section we provide some benchmarking on the Pfam-seed dataset on both token-wise and domain-wise level. We use Pfam-seed dataset splits published at [6], both the random splits and the clustered ones. The only difference is regarded to multi-domain proteins. Since we consider samples on the whole protein level rather than individual domain level, some multi-domain proteins could be mixed (contain both train and validation domains, or train and test domains etc.) thus couldn't be correctly assigned to any set. In order to keep the test set exactly the same and to keep the property of remote homology between the train, validation and test sets, such "mixed" multi-domain proteins were discarded from the train and validation sets. On the aminoacid-level (tokenwise level), our model achieved 72.6 and 95.0 % macro averaged recall on clustered and random splits respectively. In order to obtain the domain-level predictions (and to compare with the existing domain-level methods), we averaged tokenwise predictions inside the domain borders assuming they're known.

Clustered split of pfam-seed

Model	Error rate, %	Number of errors
Top Pick HMM	18.1	3844
phmmer	32.6*	6942
BLASTp	35.9*	7639
ProtCNN	27.7*	5882
ProtENN	12.2*	2590
Protomenal (ours)	20.9	4035

* the results were taken from [6].

Conclusion: The results suggest that the deep metric learning is a very promising approach for annotating multi-domain proteins, especially with few or no known homologs. The model is deployed as a web service at https://protomenal.com. We are always open for collaboration, and especially with the experimental groups interested in proteomic research.

Acknowledgements: This work was supported by the Ministry of Science and Higher Education of Russian Federation (agreement No. 075-15-2019-1665).

References

1. Eddy S. Accelerated profile HMM searches. *PLoS Comput Biol*. 2011;7:e1002195.
2. Soding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics*. 2005;21:951-960. doi: 10.1093/bioinformatics/bti125.
3. Elnaggar A. et al. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. *arXiv:2007.06225.* 2020. doi: 10.1101/2020.07.12.199554.
4. Qian Q., Shang L., Sun B., Hu J., et al. SoftTriple Loss: Deep Metric Learning Without Triplet Sampling. 2019;6449-6457. doi: 10.1109/ICCV.2019.00655.
5. Karpusha M. et al. Calibrated neighborhood aware confidence measure for deep metric learning, *arXiv:2006.04935*. 2020.
6. Bileschi M.L. et al. Using Deep Learning to Annotate the Protein Universe. *bioRxiv*. 2019;626507. doi: 10.1101/626507.