Evaluation of natural language processing embeddings in protein function prediction for bacteria

Student: Bianca-Maria Cosma[1]

B.M.Cosma@student.tudelft.nl

Supervisors: Aysun Urhan[1,2], Abigail Manson[2], Thomas Abeel[1,2]

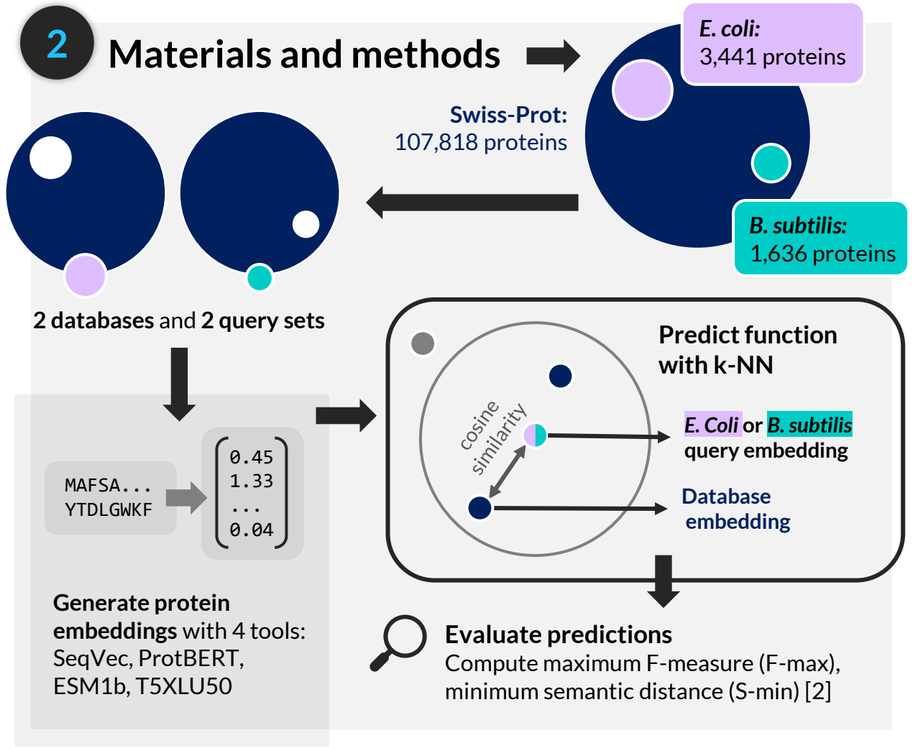
[1]Delft Bioinformatics Lab, Delft University of Technology Van Mourik, Broekmanweg 6, 2628 XE, Delft, The Netherlands;
[2]Infectious Disease and Microbiome Program, Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA, 02142, USA

1 Introduction

- Protein function prediction is the association of a protein with its role in an organism.
- Automated function prediction is a multi-label classification task with more than 45,000 labels from the Gene Ontology (GO) hierarchy [1]. GO has 3 sub-hierarchies: molecular function, biological process, and cellular component.
- Protein sequences can be represented as real-valued vectors using ideas from natural language processing embeddings.

How do unsupervised embeddings models perform in automated protein function prediction for bacteria?

2 Materials and methods



3 Results and discussion

The k-NN predictors based on embedding similarity outperformed BLAST sequence-based annotations (see Fig. 1). The model using ESM1b embeddings performed better than goPredSim [3] in all categories, and surpassed DeepGOPlus [4] for molecular function and biological process prediction.

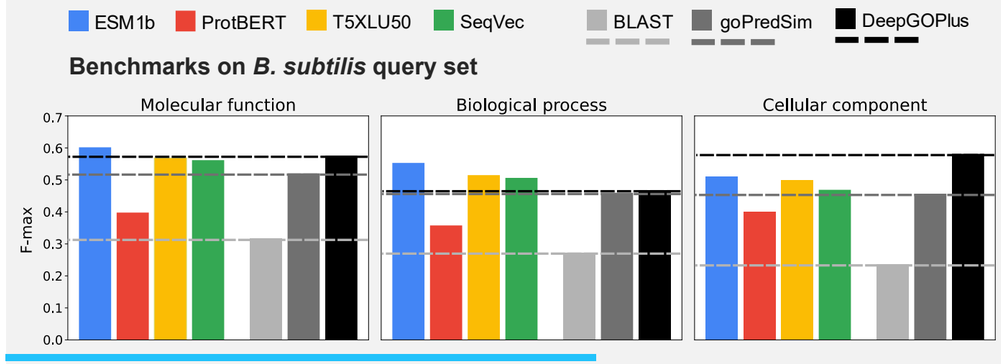


Fig. 1. The maximum F-measure (F-max) of the predictors, on the *B. subtilis* query set. Predictions from ESM1b, SeqVec, ProtBERT and T5XLU50 embeddings were made using our k-NN approach.

Previous results are reinforced by model rankings with regard to the minimum semantic distance (see Fig. 2). However, **all predictions had low recall**, which corresponds to a high rate of false negatives (see Fig. 3). This was particularly the case for *E. coli* proteins, which had more ground-truth annotations.

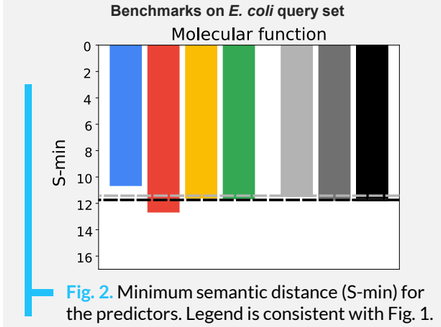


Fig. 2. Minimum semantic distance (S-min) for the predictors. Legend is consistent with Fig. 1.

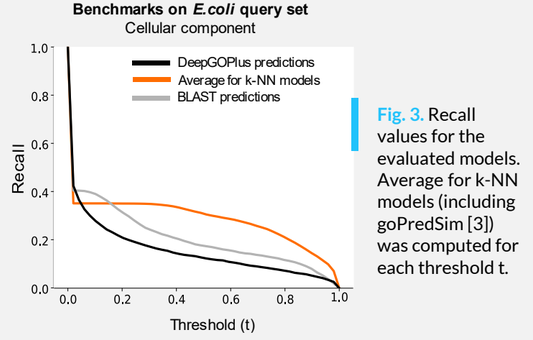


Fig. 3. Recall values for the evaluated models. Average for k-NN models (including goPredSim [3]) was computed for each threshold t.

4 Conclusion

Our k-NN models based on embedding similarity outperformed sequence-based function annotation, and their results were comparable to those from state-of-the-art predictors.

- Embeddings from deep learning can encode information about bacterial proteins beyond sequence similarity.
- The performance of all predictors was affected by high rates of false negatives.
- Complete annotation of novel bacterial protein sequences remains a prospect for future work in automated function prediction.

References

[1] Gene Ontology Consortium, "The gene ontology resource: 20 years and still GOing strong," *Nucleic Acids Research*, vol. 47, no. D1, pp. D330–D338, 2019.

[2] N. Zhou, Y. Jiang, T. R. Bergquist, et al., "The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens," *Genome Biology*, vol. 20, no. 1, p. 244, Nov. 19, 2019.

[3] M. Littmann, M. Heinzinger, C. Dallago, T. Olenyi, and B. Rost, "Embeddings from deep learning transfer go annotations beyond homology," *Scientific Reports*, vol. 11, no. 1, pp. 1–14, 2021.

[4] M. Kulmanov and R. Hoehndorf, "DeepGOPlus: Improved protein function prediction from sequence," *Bioinformatics*, vol. 36, no. 2, pp. 422–429, Jan. 15, 2020.