Comparative analysis of protein function text-based embeddings and their applicability to prediction tasks

Rohitha Ravinder [1,2], Leyla Jael Castro [1], Martin Hofmann-Apitius [2,3] and Dietrich Rebholz-Schuhmann [1,4]

[1] *ZB MED Information Centre for Life Sciences, Gleueler Str. 60, Cologne, 50931, Germany*
[2] *Bonn-Aachen International Centre for Information Technology (B-IT), University of Bonn, Friedrich-Hirzebruch-Allee 6, Bonn, 53115, Germany*
[3] *Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven 1, Sankt Augustin, 53757, Germany*
[4] *University of Cologne, Albertus-Magnus-Platz, Cologne, 50923, Germany*

Abstract

Predicting protein function is a difficult problem in bioinformatics. Many recent techniques employ embeddings to learn representations of protein sequences and infer function from these; however there have been no studies that have utilized protein function text embeddings to forecast protein function. Here, we propose to learn and explore text-driven embedding representations of protein function comment sections kept as part of the Swiss-Prot entries and understand how the resulting data can be used to enhance protein function annotations. The comparative study is based on protein function text embeddings derived from two approaches which include a combination of natural language processing frameworks such as Word2Vec, Doc2Vec and dictionary-based Named Entity Recognition and acts as a preliminary assessment based on direct propagation techniques such as sequence similarity and by-similarity prediction.

Keywords

Protein function prediction, Word embeddings, Named Entity Recognition

1. Introduction

Understanding the role of proteins is crucial to life. However, there exists only a small subset of proteins whose function is well predicted thereby making protein function prediction a fundamental task in the field of Bioinformatics. Numerous techniques have been developed for protein function prediction using sequence embeddings, protein structures or protein-protein interactions [1]. Nevertheless, to the best of our knowledge no research has been made yet that makes use of protein function text-based embeddings to evaluate their use for protein function prediction tasks. In this study, our goal is to get a better understanding of how information for protein functions can be exploited through embeddings so that the produced information can be used to improve protein function annotations.

Our work is based on the hypothesis that states a direct correlation between sequence similarity (corresponding to the BLAST identity score) and similar biological function (as expressed in the protein function comment). The idea here is to capture this correlation with the help of the corresponding protein embeddings. Here we consider the text-based embeddings that are derived from the protein function comment sections of the UniProtKB reviewed entries: SwissProt entries. Specifically, we aim to learn and compare two embedding models that map functions of protein to sequences of vector representations such that two proteins having similar function as stated in the

Proceedings Semantic Web Applications and Tools for Healthcare and Life Sciences, February 13–16, 2023, Basel, Switzerland
EMAIL: ljgarcia@zbmed.de (A. 2)
ORCID: 0000-0003-3986-0510 (A.2); 0000-0001-9012-6720 (A.3); 0000-0002-1018-0370 (A.4)

function comment section appear closer in the embedding vector space. A secondary objective is to analyze the vocabulary that emerges from the text-based embeddings.

2. Methods

Methods of representing textual data range from traditional term-frequency based methods to embeddings. In our study, we experiment and generate embeddings using two methods namely: word2doc2vec and hybrid-word2doc2vec. word2doc2vec is an approach that makes use of the Word2Vec framework [2]. This framework is a two-layer neural network that is trained to reconstruct linguistic contexts of words with each unique word being assigned to a corresponding vector using one of the two architectures: skip-gram and continuous bag-of-words. We employ a strategy to generate document embeddings from these word embeddings by calculating the centroids of all given word embeddings in a function text.

Our second method employs a hybrid approach exploring a combination of a dictionary-based Named Entity Recognition using Whatizit tool [3] and word2doc2vec framework. The dictionary-based NER approach herein aims to identify and index all the biomedical words as well as accounts for additional pieces of information such as annotations denoting Gene Ontology such as Molecular function (MF) and Biological process (BP). In order to do so, we make use of Whatizit, a text processing system based on MONQjfa, a nondeterministic and deterministic finite automata for Java. Whatizit takes as input a dictionary to recognize entities in a text and normalizes them against a controlled vocabulary. We make use of two ontologies: Medical Subject Headings (MeSH) and Gene Ontology (GO) as our controlled vocabulary for the required annotation process. These embeddings ultimately treat each function text as a document and each word in the function text as a word embedding.

3. Future Work

In order to analyze the potential offered by these embeddings we intend to perform a visualization based on clustering between both our approaches, judge clusters from protein embeddings against UniRef clusters accounting for the direct propagation technique based on sequence similarity, perform a basic analysis on the emerging protein text vocabulary as well as test prediction potential. To test the prediction potential, we intend to narrow down the emerging results to two criterias: (i) high sequence similarity but low embedding similarity and (ii) low sequence similarity but high embedding similarity. The purpose of doing so is to ultimately define and implement a strategy on how embeddings could propagate function annotations as well as estimate their scope for curation work.

4. Acknowledgements

This work was partially supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A).

5. References

[1] Villegas-Morcillo A, Makrodimitris S, van Ham RCHJ, Gomez AM, Sanchez V, Reinders MJT. Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function. Bioinformatics. 2021 Apr 19;37(2):162-170. doi: 10.1093/bioinformatics/btaa701.

[2] Mikolov T. et al. Distributed representations of words and phrases and their compositionality In: Burges C.J.C.et al. (eds) Advances in Neural Information Processing Systems. Lake Tahoe, Nevada, Vol. 26, pp. 3111–3119 (2013a).

[3] Dietrich Rebholz-Schuhmann, Miguel Arregui, Sylvain Gaudan, Harald Kirsch, Antonio Jimeno. Text processing through Web services: calling Whatizit, Bioinformatics, Volume 24, Issue 2, 15 January 2008, Pages 296–298. doi: 10.1093/bioinformatics/btm557.