GO2SUM: Generating Human readable functional summary from GO terms

Swagarika J Giri[1], Nabil Ibtehaz[1], & Daisuke Kihara[1,2,*]

1 Department of Computer Science, Purdue University, West Lafayette, IN, United States

2 Department of Biological Sciences, Purdue University, West Lafayette, IN, United States

Corresponding Author, dkihara@purdue.edu

Keywords: protein function prediction, gene ontology, language model, summarizer

Abstract

Computational protein function prediction is a crucial part of genome annotation, and the Gene Ontology (GO) database plays a vital role in identifying protein function terms. However, GO term-based protein function prediction can be challenging for those unfamiliar with the technical language and hierarchical structure of Gene Ontology. In this paper, we propose GO2Sum (Gene Ontology terms Summarizer), a novel transformer-based summarizer model that takes GO terms as input and generates a human-readable summary from predicted Gene Ontology terms. GO2Sum used a novel approach of generating a GO description document by concatenating the descriptions of annotated GO terms and feeding it into the summarizer to generate the summaries for Function CC, Subunit Substructure, and Pathway. The effectiveness of the generated summaries was assessed using six different metrics: three embedding-based (BERT, MiniLM, and BioSentVec) and three sentence-mover similarity-based approaches (WMS, SMS, and W+SMS).

Our results demonstrate that GO2Sum outperforms the vanilla T5 model for Function CC, Subunit Structure, and Pathway, with accuracy rates of 96.81%, 94.52%, and 99.27%, respectively. Furthermore, GO2Sum was also integrated into the output generated by the popular protein function prediction method Phylo-PFP.