Automated Protein Function Description for Novel Class Discovery

Meet Barot, Vladimir Gligorijevic, Kyunghyun Cho, Richard Bonneau

May 19, 2022

1 Abstract

Knowledge of protein function is necessary for understanding biological systems, but functional characterization is far outpaced by the discovery of new sequences from high-throughput sequencing technologies. Beyond the difficulty of assigning newly sequenced proteins to known functions, a more challenging issue is discovering novel protein functions. Protein function prediction, as it is usually framed in the case of Gene Ontology term prediction, is a multilabel problem with a hierarchical label space. However, this framing is limiting. It does not provide guiding principles for discovering completely novel functions. Clustering-based approaches are not able to give much information about the new functional categories that they predict; they can only predict that a protein may belong to a category that has not been studied. In this work we propose a neural machine translation model in order to generate descriptions of protein functions in natural language. The model takes as input a set of sequences, and assigns a probability to any description paired with those sequences. We define three metrics that can be computed using our model's probabilities assignment of known functional descriptions to input sequence sets. We provide quantitative results of our model in the zero-shot classification setting, scoring sequence sets with functional descriptions that the model has not seen before, as well as generated function descriptions for qualitative evaluation.

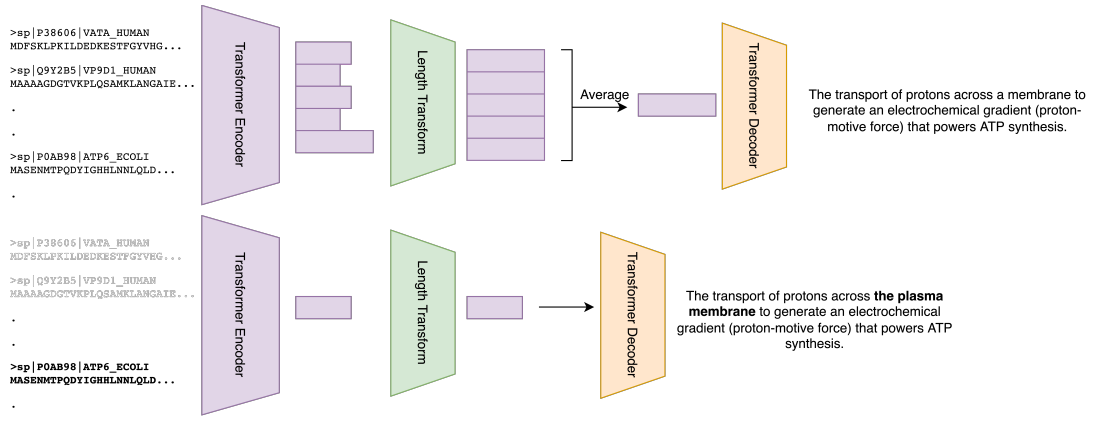

Figure 1: High-level diagram of the proposed transformer encoder-decoder model. The model is trained to produce the most specific common function of the input protein sequences.

Table 1: Sample Test Set Description Generations

True Common GO Description of Sequence Set	Model Generated Description of Sequence Set
<SOS> the process in which the anatomical structures of appendages are generated and organized . an appendage is an organ or part that is attached to the trunk of an organism . <EOS>	<SOS> the process whose specific outcome is the progression of the eye over time , from its formation to the mature structure . <EOS>
<SOS> any process that activates or increases the frequency , rate or extent of cell differentiation . <EOS>	<SOS> any process that modulates the frequency , rate or extent of cell differentiation . <EOS>
<SOS> a protein complex that contains the gins complex , cdc45p , and the heterohexameric mcm complex , and that is involved in unwinding dna during replication . <EOS>	<SOS> any process involved in forming the mature 3 ' end of a dna (mrna) molecule . <EOS>
<SOS> the targeting and directed movement of proteins into a cell or organelle . not all import involves an initial targeting event . <EOS>	<SOS> the directed movement of proteins from endoplasmic reticulum to the nucleus . <EOS>

Table 2: Model Performances. Correctness: The average number of times a correct GO term outranks an incorrect one for a given sequence set prediction. Specificity: Among correct GO terms, the average number of times a child outranks its parent. Robustness: The average rank correlation between the predictions of a pair of different identically annotated sequence sets.

Metric	Test set performance
Annotation Correctness	0.72
Specificity Preference	0.58
Annotation Robustness	0.26