Desired Attributes of a Protein Function Description Model

Meet Barot

March 9, 2022

Motivations

Why make a model that describes the common functions of a set of proteins in natural language?

1. We want to be able to predict the functions of proteins, but we are limited by the amount of data that we have in both the amount of well characterized proteins and also the variety of known functions.
2. Even the best supervised approaches can only take us to the point where we can annotate proteins that have functions that have been seen before.
3. Explicitly ontology-based zero-shot approaches such as DeepGOZero [1] do not allow for actual description of a new function that is discovered. The only information that is gained is that the protein has a new function that has some specified ontological relation to currently known functions. However, this may not sufficiently describe the new function, and it also excludes possible functions that do not directly relate to known functions.

In order to discover new categories of protein function, with some amount of information to actually design experiments to test for them, we need a model that generates functional descriptions.

The following is a list of attributes we wish the model to have.

Attribute 1: Annotation correctness.

Given a sequence set that the model is assigning scores of function descriptions:

Descriptions of GO terms that annotate the entire sequence set should be scored higher than terms that do not annotate the entire sequence set.

Let D_S be the GO term descriptions associated with sequence set S.

$$P(d \in D_S|S) > P(d \notin D_S|S)$$

A way to measure this attribute would be to calculate:

$$\frac{1}{|D_S| \cdot |D_S^c|} \sum_{d_i \in D_S, d_j \notin D_S} \mathbb{1}(P(d_i|S) > P(d_j|S))$$

where D_S^c is the complement of D_S and $\mathbb{1}$ is the indicator function.

Attribute 2: Specificity preference.

Among terms that do annotate the whole set, the model should score child terms higher than their ancestor terms. Let $A(d)$ denote the description of a direct parent of the GO term described by d.

$$P(d \in D_S|S) > P(A(d) \in D_S|S)$$

Note: any protein set that is annotated with d would always be annotated with $A(d)$, $A(A((d))$ and so on.

A way to measure this attribute would be to calculate:

$$\frac{1}{|D_S|} \sum_{d_i \in D_S} \mathbb{1}(P(d_i|S) > P(A(d_i)|S))$$

Attribute 3: Annotation robustness.

Any set of sequences that have the same exact set of GO descriptions in common should produce scores with the same rankings for those GO descriptions.

Let S_i and S_j be different sequence sets such that $D_{S_i} = D_{S_j}$ and $S_i \neq S_j$, and let $R(X)$ be a ranking function that gives the ranks of entries in X, in descending order.

$$R_d(P(d \in D_{S_i}|S_i)) = R_d(P(d \in D_{S_i}|S_j))$$

A way to measure this attribute would be to calculate the average Spearman's rank correlation of the rankings for all sequence sets' correct descriptions. Let $R_{S_i} = R(P(D_{S_i}|S_i))$:

$$\frac{1}{N \cdot (N-1)} \sum_{S_i, S_j} \frac{\text{cov}(R_{S_i}, R_{S_j})}{\sigma_{R_{S_i}} \sigma_{R_{S_j}}}$$

where N is the total number of sequence sets that have the exact set of GO descriptions D_{S_i}. In reality, this number may be too large to actually sum (especially if $|D_{S_i}|$ is small), so we would approximate this measure by subsampling $n < N$ sequence sets to average over instead. The sum is only calculated over non-identical pairs of sequence sets.

Additional Evaluation

As these scoring metrics for evaluation are automated, they can be used for optimizing the architecture and other hyperparameters of the model (either manually or with some search method). However, in the case of actual use on proteins that are not very well studied, it can be difficult to know whether a given description is accurate. Human-assisted evaluation will be needed for the descriptions generated for a given set of novel proteins. This feedback could be used to fine-tune the model to produce more accurate, fluid or generally desirable descriptions of proteins, as has been done for document summarization models [3, 2].

One possible way of obtaining human feedback would be to ask an expert with knowledge of the Gene Ontology and familiarity with some families of proteins to choose between two descriptions for a given sequence set that is generated from a trained model.

Doing this over a large enough dataset would allow us to train a reward estimation model that can then be used to fine-tune the original trained model using reinforcement learning. However, this would be expensive, as the task needs to be done by an expert. Richer information, like ranking the similarities to an existing GO term or suggesting changes to particular portions of the description could be used to get feedback.

References

[1] Maxat Kulmanov and Robert Hoehndorf. "DeepGOZero: Improving protein function prediction from sequence and zero-shot learning based on ontology axioms." In: *bioRxiv* (2022). DOI: 10.1101/2022.01.14.476325. eprint: https://www.biorxiv.org/content/early/2022/01/14/2022.01.14.476325.full.pdf. URL: https://www.biorxiv.org/content/early/2022/01/14/2022.01.14.476325.

[2] Nisan Stiennon et al. "Learning to summarize with human feedback." In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3008–3021.

[3] Daniel M Ziegler et al. "Fine-tuning language models from human preferences." In: *arXiv preprint arXiv:1909.08593* (2019).