Desired Attributes of a Protein Function Description Model

Meet Barot

February 28, 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:

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

where D_S^c is the complement of D_S.

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:

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

Attribute 3: Annotation robustness.

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

Let S_i and S_j be different sequence sets such that $D_{S_i} = D_{S_j}$.

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

A way to measure this attribute would be to calculate:

$$\frac{1}{N^2 \cdot |D_{S_i}|} \sum_{S_i, S_j, d_k \in D_{S_i}} |P(d_k|S_i) - P(d_k|S_j)|$$

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.

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.