Desired Attributes of a Protein Function Description Model

Meet Barot

February 25, 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 could be gained is that the protein has a new function that has some specified ontological relation to currently known functions, but this may not sufficiently describe this new function, and it also excludes possible functions that do not directly relate to known functions.

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

Attribute 1: Annotation coherence.

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.

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

Attribute 2: Specificity preference.

Among terms that do annotate the whole set, the model should score more specific terms higher than less specific terms.

Let $t(d)$ be the depth of a GO term description d, and δ is an arbitrary depth.

$$P(d \in D_S, t(d) \geq \delta | S) > P(d \in D_S, t(d) < \delta | S)$$

Attribute 3: Branch equality.

Among the most specific terms that annotate the whole set, descriptions from all three branches of GO should be scored equally.

Let B_i be the ith GO branch, and D_{S,B_i} be the descriptions associated with sequence set S that are in branch B_i.

$$P(d \in D_{S,B_i}, t(d) = \delta | S) = P(d \in D_{S,B_j}, t(d) = \delta | S)$$

Attribute 4: 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 $D_{S_i} = D_{S_j}$.

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

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.